Experimental Methodology and its Applications in Economics

by

Garrett M. Petersen

M.A., Queen's University, 2013

B.Sc., University of Victoria, 2012

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in the

Department of Economics

Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Spring 2021

Declaration of Committee

Name: **Garrett M. Petersen**

Degree: **Doctor of Philosophy**

Thesis title: **Experimental Methodology and its Applications in Economics**

Committee:

Chair: Alexander Karaivanov
Professor, Economics

David Freeman
Supervisor
Associate Professor, Economics

Douglas W. Allen
Committee Member
Professor, Economics

Shih En Lu
Examiner
Associate Professor, Economics

Michael Makowsky
External Examiner
Associate Professor, Economics
Clemson University

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

This dissertation explores and applies experimental methods in economics. The first two chapters deal with the methodology of lab experiments, while the third presents a study on mobility apps. In the first chapter, I examine deliberating groups in a jury-like setting where subjects have private information and an opportunity to discuss it before a vote. The study uses a belief elicitation mechanism to incentivize subjects to truthfully report their beliefs both before and after they deliberate, allowing for the measurement of the change in beliefs. I find that deliberation tends to reduce the average error in beliefs, measured as the difference between the belief and the true outcome. The basic experiment follows past deliberation experiments in the literature. It features an abstract setting with private signals in the form of a randomly drawn red or blue ball. To test whether the results are generalizable, I replicated this experiment in a framed setting where subjects read the evidence from a real murder trial. I found no difference between the results of the experiments in these two different settings. The second chapter investigates the use of reinforcement methods in lab experiment instructions. We experimentally compare how methods of delivering and reinforcing experiment instructions impact subjects' comprehension and retention of payoff-relevant information. We find combinations of reinforcement methods that can eliminate half of non money-maximizing behaviour, and we find that we can induce a similar reduction via enhancements to the content of instructions. Residual non money-maximizing behaviour suggests this may be an important source of noise in experimental studies. The third chapter diverges from lab experiments to study Mobility as a Service (MaaS). We test whether a multimodal route-planning service caused users to use combined routes featuring both ride hailing and transit. We find that ride-hailing trips connected with rail stops increased from 3.0% of trips to 5.5% among existing users. In areas where the feature supported bus connections, trips connecting to bus stops increased from 4.6% to 8.7% among existing users.

Keywords: Lab experiments, deliberative voting, mobility as a service.

Dedication

For Calvin.

Acknowledgements

Thank you to my wife, Danika, and to my family. I would not be here without your support.

Thank you to all the professors, colleagues, and friends who taught me to think like an economist.

Table of Contents

List of Tables

List of Figures

Chapter 1

Three Angry Men: A Framed Jury Experiment

Garrett M. Petersen[†]

Abstract

Many institutions rely on deliberating groups to reach decisions in situations with limited information. In this study, I examine deliberating groups in a jury-like setting where subjects have private information and an opportunity to discuss it before a vote. The study uses a belief elicitation mechanism to incentivize subjects to truthfully report their beliefs both before and after they deliberate, allowing for the measurement of the change in beliefs. Deliberation tends to reduce the average error in beliefs, where error is measured as the difference between a reported belief and the truth. The basic experiment featured an abstract setting with private signals in the form of a randomly drawn red or blue ball. To test for generalizability, I replicated this experiment in a framed setting where subjects read the evidence from a real murder trial. I found no difference between the results of the experiments in these two different settings.

Keywords: Jury trials, deliberative voting, belief updating, information aggregation.

JEL Classification: C92, D70, D80, K40

Deliberation is a decision-making process that involves a period of discussion followed by a vote. Government legislatures, central bank committees, corporate boards, and juries in both civil and criminal courts all use deliberation to reach decisions. Deliberation

[†]I am grateful to my supervisor, David Freeman, for very helpful feedback and guidance. I would also like to thank Erik Kimbrough, Alexander Billy, Douglas Allen, and the seminar audiences at SFU, the Public Choice Society annual meeting, and the Canadian Economics Association annual meeting for their feedback and comments.

allows more information aggregation than simple voting, as a deliberating group can share private information before reaching a decision. Each individual can make a more informed decision about how to vote by incorporating others' private information. In principle, this means that deliberating groups can make better decisions than they would by voting without deliberating. However, this benefit depends on the members truthfully sharing their information and efficiently incorporating it into their votes, which is far from guaranteed.

The existing theoretical and experimental literature on deliberation discusses the practice in an abstract setting. In the canonical deliberation experiment, there are two possible states of the world (red and blue) and each subject is shown a private signal (in the form of a red or blue ball) with a known probability of matching the true state (Guarnaschelli et al., 2000; Goeree and Yariv, 2011; Bhattacharya et al., 2014; Bouton et al., 2017). Subjects communicate and then vote according to a decision rule that varies from experiment to experiment. This framework allows the researcher to induce subjects to hold certain beliefs by delivering a signal that can only imply one possible belief. For instance, if the researcher informs subjects that there is always a 70% chance that a signal matches the true state, a red signal can only imply a 70% chance of red. After receiving the signal, subjects may choose to share their signals with each other before voting. The outcome of the vote determines the final outcome according to the decision rule. This framework was designed to correspond closely to theoretical models of deliberating groups.

Although beliefs are important at every stage of the deliberation process, researchers cannot observe them. As a result, past experiments have informed subjects of the distribution of possible signals so that a subject receiving a given signal has enough information to compute the probability of a given outcome. That subjects actually compute this probability is a modelling assumption that does not need to be literally true for deliberation models to deliver valuable insights. The experimental setup used throughout the literature has the advantage of providing researchers with a high level of control, but it raises questions about whether real-world deliberation is similar enough to laboratory deliberation for the results to hold across settings.

A core assumption of all the experiments in the experimental deliberation literature is that deliberation occurs in a similar way in different environments (e.g. in the laboratory and in a real jury). This study tests that assumption by modifying the standard experimental design to allow for subjectivity in the signals presented to subjects. The standard deliberation experiment features a private signal, a communication round, and a vote, in that order. All treatments in this study include an additional voting round before the communication round and a belief elicitation during each of the two votes. This change has two advantages. First, beliefs do not need to be imposed if they can instead be measured. Second, within-subject changes in votes and beliefs can be observed by comparing subjects' decisions before and after deliberation.

This study replicated the standard deliberation experiment described above as a baseline, including the additional voting round and elicitations. In this treatment, the true state was either red or blue, and each subject was shown an independent draw of a red or blue ball. This ball had a 70% chance of matching the true state. A second treatment included a subjective task where subjects observed a grid of 400 red and blue balls and attempted to judge which colour occurred more frequently. Finally, the study included a framed treatment where subjects read news reports from real jury trials and attempted to determine the guilt of the defendant.

The experiment produced three key results. First, deliberation successfully aggregates individuals' private information. The average belief error fell from 44.6% before deliberation to 38.4% afterwards. Second, for a given set of initial beliefs in a group, deliberation leads people to update their beliefs in a similar way regardless of the informational setting. A regression model of belief updating did not find significant differences between the treatments. This result provides evidence o the external validity of deliberation experiments in abstract laboratory settings.

Finally, a supplementary benefit of measuring both beliefs and votes is that comparing them makes it possible to infer the threshold at which subjects change their votes. Theory predicts that jurors who believe they may have conflicting voting thresholds have some incentive to not communicate truthfully (Austen-Smith and Feddersen, 2006). The third key result is that most jurors vote consistently with a voting threshold of 50%, even if their incentives suggest a different threshold. This is consistent with the heuristic strategy observed by Le Quement and Marcin (2020), wherein people share their signals truthfully and then vote with the majority of signals regardless of strategic concerns.

The study of deliberative decision making, wherein a group can share information before a vote, grew out of the literature on strategic voting in the Condorcet jury model framework (Feddersen and Pesendorfer, 1996, 1997, 1998). In Condorcet jury models, a group of people with aligned interests but limited information must vote on a binary decision. The central idea in the strategic voting literature is that if voters care about the vote's outcome, each must vote under the assumption that their vote is pivotal. Under certain decision rules, this can mean voting against one's private information. The theory of deliberative voting was developed by adding communication prior to the voting process. Coughlan (2000) and Gerardi and Yariv (2007) find that jurors with sufficiently aligned interests will truthfully share their private information before voting, making that private information public and thereby eliminating differences between voting rules.

Following Coughlan's (2000) theoretical framework, Guarnaschelli et al. (2000) developed the standard experimental design that characterizes the experimental deliberation literature. In their jury experiment, Guarnaschelli et al. varied the group size (3 or 6), decision rule (majority or unanimity), and presence of a straw poll. Their findings were consistent with deliberation theory. In the absence of communication, jurors show signs

of strategic voting. However, communication leads most jurors to truthfully reveal their signals and vote with the majority of signals. Goeree and Yariv (2011) conducted a similar experiment with unrestricted communication replacing the straw poll. They introduced a treatment with asymmetric and private incentives, theorized by Austen-Smith and Feddersen (2006) to prevent truthful information sharing under a unanimity rule. Contrary to theory, Goeree and Yariv showed that jurors tend to over-share information truthfully despite incentive asymmetry.

Fehrler and Hughes (2018) modified this framework to model committees subject to professional consequences from outside observers. They assigned each juror to be either a high-information or low-information type, where there are negative consequences to publicly revealing that one is a low-information type. The authors found that making deliberation transparent to outside observers negatively affects the group's ability to aggregate their private information.

By testing subjects' preferences before a jury experiment, Le Quement and Marcin (2020) tested various theories to explain why subjects tend to be truthful when doing so does not maximize their expected payoffs. Le Quement and Marcin found that roughly 80% of subjects play heuristically, sharing their signals truthfully and then voting with the majority signal, while the other 20% play more strategically, sharing information and voting in a way that approximates theoretical predictions.

1.1 Theory

The theory in this section loosely follows the setup in Austen-Smith and Feddersen (2006), with the slight modification that the signal strength is not consistent across jurors. In the standard Condorcet jury model, a jury comprised of three people must decide whether to convict or acquit a criminal defendant. The guilt or innocence of the defendant is assigned randomly with equal probability before the jurors observe the evidence. Each juror receives a private and independent signal related to the guilt of the defendant. Denote juror i's signal as $s_i \in (0, 1)$. Jurors can report their signals to each other prior to voting, either truthfully or not. For simplicity, I assume the decision of what signal to report is simultaneous. Finally, jurors vote for conviction or acquittal, with the majority vote determining the outcome.

The Signal Structure The odds of receiving a signal of s_i is s_i if the defendant is guilty and $1 - s_i$ if the defendant is innocent. This means that $P(Guilty|s_i) = s_i$. Furthermore, multiple signals can be combined according to Bayes' rule. The likelihood-ratio form of the rule takes the following form given independent signals:

$$\frac{P(\text{Guilty}|s_1, \ldots, s_n)}{P(\text{Innocent}|s_1, \ldots, s_n)} = \prod_{i=1}^{n} \frac{s_i}{1 - s_i} \tag{1.1}$$

The combination of signals depends on whether other jurors share them.

1.1.1 behavioural Model

Jurors' preferences are solely related to the trial's outcome (acquittal or conviction) and the true state of the world (innocence or guilt). Assume symmetry of preferences for now: Jurors get a high payoff of H if a correct outcome occurs (convicting the guilty or acquitting the innocent) and a low payoff of L if an incorrect outcome occurs (convicting the innocent or acquitting the guilty). Let $H > L$, and assume common knowledge of these preferences.

The expected utility, EU, for all jurors given the probability of guilt is

$$EU = \begin{cases} P(\text{Guilty})H + P(\text{Innocent})L \text{ for Conviction} \\ P(\text{Guilty})L + P(\text{Innocent})H \text{ for Acquittal.} \end{cases} \tag{1.2}$$

It follows from Equation (1.2) that when P(Guilty)=P(Innocent)=0.5, jurors are indifferent between conviction and acquittal. If P(Guilty) is higher than 0.5, jurors prefer to convict, while if it is lower than 0.5 they prefer to acquit.

Given this relationship between jurors' beliefs about guilt and their votes, is it optimal for jurors to truthfully share their signals with their fellow jurors? Since jurors only care about achieving the correct outcome, they will report a signal that maximizes the odds of their fellow jurors voting correctly.

Strategic voting implies that jurors should vote as if they were the pivotal voter since this is the only case in which their vote determines the outcome.[1] In this particular setup (majority voting with three jurors and a flat prior about the superior outcome), being the pivotal voter means that the other two voters voted differently from one another. Since these votes provide equal and opposite signals about the true state (guilt or innocence), the pivotal vote should vote in accordance with their own signal.[2]

Let $\tilde{s}_i$ be the signal reported by juror i. Assuming juror j believes that juror i is reporting the signal truthfully, j's updated belief will be

$$\frac{P(\text{Guilty}|s_i, s_j)}{P(\text{Innocent}|s_i, s_j)} = \frac{s_i}{1 - s_i}\frac{s_j}{1 - s_j}. \tag{1.3}$$

For $\tilde{s}_i > 0.5$, juror j's vote will be changed from acquittal to conviction so long as $1 - \tilde{s}_i < s_j < 0.5$. Similarly, for $\tilde{s}_i < 0.5$, juror j's vote will be changed from conviction to acquittal so long as $0.5 < s_j < 1 - \tilde{s}_i$.

[1] This follows from the assumption that jurors are only motivated by the outcome of the vote. This may not always be true in practice. For instance, it may not hold when people engage in expressive voting. (see Tyran, 2004).

[2] The strategic vote is also a naïve vote in this scenario, since the assumption that one's vote is pivotal doesn't give extra information.

It follows that $\tilde{s}_i = s_i$ is the optimal reported signal. Given a truthfully reported s_i, juror j will vote to convict if $P(\text{Guilty}|s_i, s_j) > 0.5$ and to acquit if $P(\text{Guilty}|s_i, s_j) < 0.5$. This is optimal from juror i's perspective; if they reported $\tilde{s}_i > s_i$ then juror j would sometimes vote for conviction when $P(\text{Guilty}|s_i, s_j) < 0.5$. Similarly, if he reported $\tilde{s}_i < s_i$ then juror j would sometimes vote for acquittal when $P(\text{Guilty}|s_i, s_j) > 0.5$. The same is true of the third juror.

Since being truthful is an optimal strategy if others assume you are truthful, and assuming others are truthful is an optimal strategy if they really are truthful, the truthful reporting of signals is an equilibrium. Therefore, all jurors truthfully report their signals and vote unanimously for the outcome that is more likely to deliver a high payoff given all the signals.

Partisan Incentives Sometimes, real jurors enter the courtroom with preconceptions that affect how they vote. To model this, suppose that all jurors have heterogeneous and private incentives. Assume that each juror has an equal (and independent) probability of being a partisan for conviction or acquittal and that no jurors are non-partisan. A partisan receives a payoff of $3H$ if the jury correctly chooses his partisan outcome. These incentives match those in the "weak partisan" treatment of Goeree and Yariv (2011).

$$EU_C = \begin{cases} P(\text{Guilty})3H + P(\text{Innocent})L \text{ for Conviction} \\ P(\text{Guilty})L + P(\text{Innocent})H \text{ for Acquittal.} \end{cases} \tag{1.4}$$

$$EU_A = \begin{cases} P(\text{Guilty})H + P(\text{Innocent})L \text{ for Conviction} \\ P(\text{Guilty})L + P(\text{Innocent})3H \text{ for Acquittal.} \end{cases} \tag{1.5}$$

Equations (1.4) and (1.5) show the modified payoffs, where EU_C and EU_A are the expected utility of partisanship for conviction and acquittal, respectively. It follows that a partisan for conviction is indifferent between conviction and acquittal when $P(\text{Guilty}) = 0.25$ and a partisan for acquittal is indifferent when $P(\text{Guilty}) = 0.75$. In other words, all jurors vote for acquittal when $P(\text{Guilty}) < 0.25$, for conviction when $P(\text{Guilty}) > 0.75$, and for their partisan bias when $0.25 < P(\text{Guilty}) < 0.75$.

Truthfully reporting one's signal is not generally an equilibrium in this scenario. Suppose that Jurors 1 and 2 are both truthful and credulous. Juror 3 is a conviction partisan and is deciding whether to truthfully report his signal. Suppose that he is able to first observe the signals reported by Jurors 1 and 2 before sending his own.[3] He observes the signals reported

[3]If Juror 3 could not observe the other jurors' signals prior to reporting his own, his problem would be considerably more complex. Reporting a higher signal to mislead those with the opposite bias runs the risk of misleading those with the same bias, and the juror would need to trade off these competing possibilities against each other.

by Jurors 1 and 2 and realizes that $P(\text{Guilty}|s_1, s_2, s_3) \in (0.25, 0.75)$. If Juror 3 truthfully reports s_3, then all jurors will vote for their partisan bias. This means there is a 25% chance of an acquittal, since Jurors 1 and 2 each have a 50% chance of being acquittal partisans.

If Juror 3 reports a sufficiently high signal such that $P(\text{Guilty}|s_1, s_2, \tilde{s}_3) > 0.75$, he can guarantee the other two jurors will vote for conviction, raising his expected payoff. This violates the assumption of truthfulness and gives the other jurors an incentive to disbelieve the signals that Juror 3 reports. Communication breaks down when preferences are heterogeneous and uncertain.

1.1.2 Predictions

This simple model of jury behaviour implies several predictions about votes and beliefs and how they are affected by deliberation. In the experiment, I observed jurors' beliefs and votes before and after deliberation in a variety of settings under both aligned and partisan incentives. The following hypotheses are testable claims implied by the model.

I assume that jurors form beliefs in accordance with the laws of probability (i.e. Bayes' rule). They were given a known prior probability (always 50% by design) and a signal they could use to update their belief. When that signal was objective and drawn from a known probability distribution, this implied a specific posterior belief.

Hypothesis 1. *When the evidence permits jurors to arrive at a mathematically correct belief about the world, they will report that belief.*

Jurors should be able to distinguish relevant information that is part of their signal from irrelevant information that has no relationship to the defendant's guilt. The experimental setup included both a meaningful signal and orthogonal information about jurors' own payoffs. The model implies that information on the payoff structure should not affect jurors' beliefs.

Hypothesis 2. *Irrelevant information does not affect jurors' beliefs.*

The model assumes that jurors will vote according to their beliefs. Given symmetric incentives, jurors simply vote according to the outcome they believe is more likely, regardless of their risk preferences. When incentives are asymmetric (i.e. partisan), risk preferences can become relevant. Hypothesis 3 assumes risk neutrality over the small stakes of the experiment.[4]

Hypothesis 3. *Jurors vote for the outcome that maximizes their expected payoff given their beliefs.*

[4]Rabin (2000) shows that risk aversion over small stakes implies bizarre preferences over large ones. This study had small stakes, so one should expect rational people to be risk-neutral. While previous laboratory experiments have indicated that people tend to be risk-averse even over small stakes (Holt and Laury, 2002, 2005), I use risk neutrality as an initial assumption. I discuss alternative assumptions in Section 1.3.

The model implies that jurors will share their signals truthfully absent partisan incentives. Moreover, they will update their beliefs to align more closely with those of the group. Their beliefs will become more accurate (on average) as a result.

Hypothesis 4. *When jurors' incentives are aligned, deliberation causes their beliefs to update toward those of other jurors. This makes the jurors' beliefs more accurate on average.*

The model predicts that communication will break down when jurors can have conflicting partisan incentives. Partisan incentives generate an incentive to lie. Jurors can be expected to recognize this incentive to lie, and so they will not believe one another.

Hypothesis 5. *When incentives are not aligned, jurors do not update their beliefs in response to deliberation.*

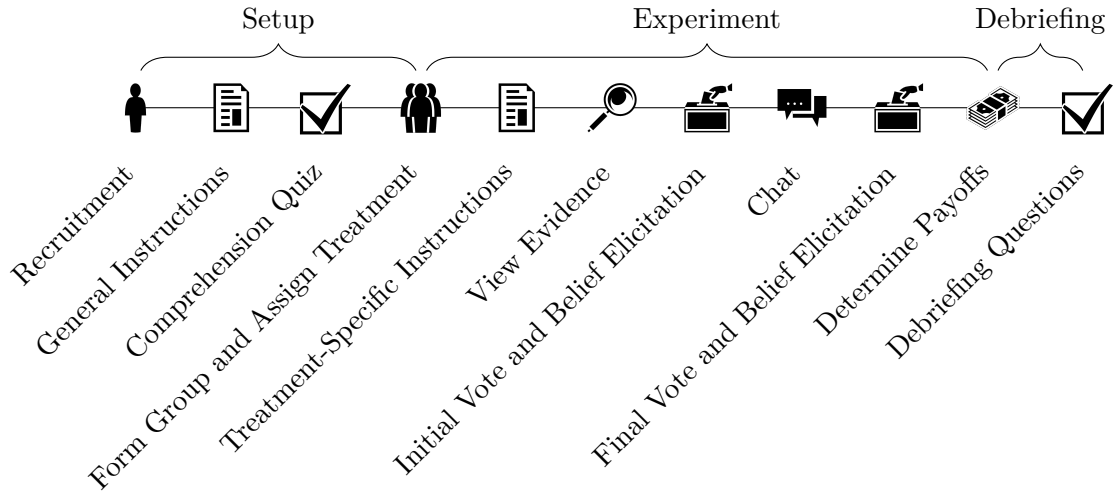
In order to generalize from theory and laboratory experiments to real-world settings, the deliberation process must be sufficiently consistent across settings to generate comparable outcomes. One key difference between past laboratory experiments and the real world is how people absorb information and form beliefs. The model assumes that the form of signals is neutral to the deliberation process.

Hypothesis 6. *The deliberation process is independent of the form of the signals presented to jurors.*

1.2 Experimental Method

In this section, I discuss the experiment itself, leaving the details of the recruitment, interface, setup, and debriefing to Section 1.2.3. Figure 1.1 shows the course of the experiment.

Figure 1.1: Experimental Timeline



The experimental setup followed the institutional framework established in Section 1.1, with the addition of a pre-deliberation voting round used to establish the effect of deliberation. Three subjects were grouped together to form an experimental jury, and this jury was assigned randomly to a treatment. All treatments followed the same general process. First, the true state of the world was randomly determined. This true state can be conceptualized as the actual guilt or innocence of a criminal defendant. Next, each juror viewed evidence relating to this true state. Once the jurors had assessed this evidence, and prior to any deliberation, they conducted an initial vote. Each juror individually stated their belief about the true state as a percentage using a slider from 0% to 100%. Next, the jury entered the deliberation phase, communicating freely through a chatbox for three minutes. After deliberation, the jurors conducted a second vote and had a second chance to report their beliefs.

Next, the computer randomly selected one of the four decisions for payment: initial vote, initial belief, final vote, or final belief. If a vote was selected for payment, the participants received a high payment ($0.50–1.50 depending on the treatment, explained in Section 1.2.1) if at least two out of the three participants voted correctly (i.e. to convict the guilty or acquit the innocent). They received a low payment ($0.10) if at least two voted incorrectly. If a belief report was selected for payment rather than a vote, payment was based on the Karni (2009) mechanism, discussed in Section 1.2.2 below.

1.2.1 Treatments

The experiment included five treatments: objective nonpartisan, objective partisan, subjective nonpartisan, subjective partisan, and framed. Each group was randomly assigned to one treatment. The treatments affected the form of the evidence delivered to jurors (objective, subjective, or framed) and the incentive structure (nonpartisan or partisan).

Objective Nonpartisan Treatment In this treatment, the true state of the world was the colour of an unseen jar. The jurors were informed that the red jar (corresponding to guilt) contained seven red balls and three blue ones and that the blue jar (corresponding to innocence) contained seven blue balls and three red ones. Evidence was presented as a single draw from the jar, with replacement. All three members of a group were shown independent draws. If most group members voted correctly, all members received the higher payoff of $1.00.

Objective Partisan Treatment This treatment is identical to the previous one except for the high payoff. At the start of the experiment, each juror in a group was randomly and independently assigned as a red or blue partisan with equal probability. If a vote was selected for payment, and the majority correctly voted for red, red partisans received $1.50, and blue partisans received $0.50. These payoffs were reversed if the majority correctly

voted for blue, with blue partisans getting $1.50 and red partisans getting $0.50. As in the other treatments, everyone received the same low payoff of $0.10 if a majority of the group voted incorrectly, regardless of their partisan identity.

Subjective Nonpartisan Treatment This treatment adapts an experimental task used in Caplin and Dean (2015) and Magnani and Oprea (2017). After the true state of the world was determined (red or blue), jurors were shown a 20×20 grid of red and blue balls. If the true state of the world was red, the grid contained 205 red balls and 195 blue balls. If the true state of the world was blue, the numbers were reversed. All jurors in a group were shown the same grid and given 10 seconds to look at it. The payoffs were identical to those in the objective, nonpartisan treatment: $1 if a vote was selected and the group voted correctly, regardless of colour.

Subjective Partisan Treatment This treatment combines the grid of balls from the subjective nonpartisan treatment with the partisan incentives from the objective partisan treatment. Before viewing the grid, jurors were informed of whether they were a red or blue partisan and how their payoffs differed as a result. Partisan incentives were identical to those in the objective partisan treatment: $1.50 if the group correctly voted for the outcome matching a juror's partisan leaning, $0.50 if the group correctly voted for the outcome that did not match the juror's partisan leaning, and $0.10 if the group voted incorrectly either way.

Framed Treatment The framed treatment was designed to achieve a higher level of realism than a traditional experiment. Jurors read the evidence from an actual murder trial as reported by a journalist who was present in the courtroom. The jurors had up to 25 minutes to read the evidence, or until all members of the group were finished. As before, the true state of the world was randomly determined, with an equal probability of a guilty or innocent defendant. This true state of the world determined which trial the group read about: one selected from the National Registry of Exonerations (innocent) or one that resulted in a conviction that was never overturned (guilty).[5] The deliberation time in this treatment was extended from three minutes to five to allow for more discussion.

The payoffs in the framed treatment were the same as in both the nonpartisan treatments, but with an extra $0.50 bonus for all participants regardless of the outcome to compensate them for the extra time.

[5]The details of the selection process for real trials are discussed in Appendix A.1.

1.2.2 Payoffs and Elicitation

If a belief was selected for payment, the payoffs were determined according to a belief elicitation based on the Karni (2009) mechanism.[6] The mechanism worked in the following way. Each juror was randomly assigned a percentage from 0% to 100%. Jurors did not know their own percentages, only that all values from 0% to 100% were equally likely. If their reported belief in guilt was greater than or equal to their randomly assigned percentage, the juror received \$1 if the defendant was guilty and \$0.10 if the defendant was innocent. Otherwise, they received \$1 with a random probability equal to their randomly assigned percentage and \$0.10 otherwise. By reporting a belief of say, 63%, a juror expressed that they were indifferent between a bet on the defendant's guilt and a bet they would win with 63% probability. Thus, this mechanism incentivized jurors to truthfully report their beliefs to maximize their odds of a high payoff.

1.2.3 Recruitment and Online Interface

The experiment was coded in o-Tree (Chen et al., 2016), and the subjects were recruited via Amazon's Mechanical Turk (MTurk) through the TurkPrime platform (Litman et al., 2017). The sample was limited to MTurk users in the United States. Ninety-six groups of three participants successfully completed the experiment between November 2018 and March 2019, for a total of 288 participants. All payments were made in USD. Each subject was paid \$2 for completing the experiment in addition to the variable payment determined by their performance in the experiment.

I opted for MTurk over a laboratory experiment because MTurk allows for larger sample sizes at a lower cost (Buhrmester et al., 2011). While some have questioned the reliability of MTurk for experimental research, Snowberg and Yariv (2018) showed that MTurk users behave similarly to both student populations and a representative sample of Americans. The study was designed with multiple checks to select attentive subjects and ensure high-quality responses. Study participation was limited to MTurk users who had completed at least 100 Human Intelligence Tasks (HITs) on MTurk with at least a 98% approval rating on those tasks. Participants joined the experiment through the web interface, read through the general (non-treatment-specific) instructions, and completed a quiz to test whether they had read and understood the instructions. If a subject answered a question incorrectly, they were informed that their answer was wrong and allowed to try again after reviewing the instructions more carefully. Each participant could make a maximum of 20 mistakes across all attempts before they were prevented from proceeding. As a result, a respondent who was answering randomly would be prevented from entering the experiment, while an attentive person could easily pass the quiz. Once they had answered all the quiz questions correctly,

[6] This was presented to participants using a narrative about robots borrowed from Coffman (2014).

the participants were sent to a virtual waiting room until they could be paired with two other participants to form an experimental jury. These practices build on and exceed the best practices used by other researchers to ensure engaged responses from study participants recruited through MTurk (Kennedy et al., 2018).

1.3 Analysis and Results

In this section, I analyze the results of the experiment and revisit the hypotheses laid out in Section 1.1.2 to assess whether they fit the data.

Table 1.1: Summary statistics

	Objective nonpartisan	Objective partisan	Subjective nonpartisan	Subjective partisan	Framed
Participants	72	45	48	39	84
Groups	24	15	16	13	28
Initial correct votes	45 (62.5%)	34 (75.6%)	31 (64.6%)	20 (51.3%)	49 (58.3%)
Final correct votes	47 (65.3%)	37 (82.2%)	32 (66.7%)	26 (66.7%)	57 (67.9%)
Participation fee	$2.00	$2.00	$2.00	$2.00	$2.50
Average variable payoff	$0.59	$0.82	$0.61	$0.81	$0.59
Average total earnings	$2.59	$2.82	$2.61	$2.81	$3.09

Table 1.1 summarizes the size of each treatment, along with the average payoffs and voting patterns. The remaining results follow the flow of the experiment: first, I discuss the results related to jurors' beliefs, then the results related to their voting behaviour, and then the results related to deliberation.

1.3.1 Beliefs

In the first step of the experiment, participants viewed evidence and formed beliefs.

Result 1. *Jurors are not generally good at forming the mathematically correct beliefs implied by their signals.*

In the objective treatments, the signals implied specific probabilities. Jurors were informed that one of two jars would be selected with equal probability. They were also told that the red jar contained seven red balls and three blue balls, while the blue jar contained seven blue balls and three red ones. If this were a mathematics test rather than an experiment, there would be one correct answer for a juror presented with a red ball: a 70% chance of a red jar. Similarly, a blue ball yields a 30% chance of a red jar.

Table 1.2: Actual beliefs versus mathematically correct ones

	Before Deliberation		After Deliberation			
Red balls:blue balls	1:0	0:1	3:0	2:1	1:2	0:3
Probability of red jar	0.7	0.3	0.93	0.7	0.3	0.07
Exactly correct	11	10	0	3	5	0
Correct +/- 0.01	11	10	0	3	5	0
Correct +/- 0.05	14	26	1	5	17	4
Correct +/- 0.10	30	29	7	11	22	8
Correct side of 0.50	38	47	17	20	33	22
n	50	67	18	24	48	27

As Table 1.2 shows, most people did not get this correct, falsifying Hypothesis 1. Only 21 out of 117 (17.9%) jurors in these treatments reported a mathematically correct initial belief, while 59 (50.4%) fell within 10 percentage points. Thirty-two (27.4%) jurors reported initial beliefs that failed to assign a higher probability to the colour they were shown, demonstrating considerable confusion. As for post-deliberation beliefs, only eight (6.8%) jurors reported the exact mathematically correct beliefs given their group's signals after deliberation. Only 25 (21.4%) jurors assigned a higher probability to the colour that had been drawn less in their group.

Fewer jurors reported mathematically precise and correct answers after deliberation, likely owing to the greater difficulty of calculating probabilities from multiple independent draws. However, there were also fewer severely confused jurors (whose beliefs fell on the wrong side of 50%), perhaps owing to guidance from the others in their group.

Result 2. *Jurors' beliefs can be swayed by irrelevant information.*

Table 1.3: Effect of partisan bias on beliefs

	Dependent variable:	
	Initial belief	Final belief
	(1)	(2)
Partisan toward conviction	0.162^{***} (0.054)	0.049 (0.063)
Constant	0.413^{***} (0.040)	0.504^{***} (0.046)
Observations	84	84

*p<0.1; **p<0.05; ***p<0.01

Model (1) of Table 1.3 shows how initial beliefs are skewed towards jurors' partisan leanings, falsifying Hypothesis 2. If jurors were forming their beliefs in the way implied by Hypothesis 2, based on their signal and ignoring irrelevant information, then the coefficients in Model (1) should be

$$\text{Initial belief} = 0(\text{Partisan toward conviction}) + 0.5 + \varepsilon \tag{1.6}$$

where ε captures the effect of the signal and is independent of the other terms. Instead, there was a constant of 0.413, which differs significantly from 0.5 ($p = 0.016$), and a positive significant coefficient on partisanship of 0.162. The average guilt partisan had a mean initial belief of 57.6%, while the average acquittal partisan had a mean initial belief of 41.3%.[7] This is surprising, since participants were informed that partisan leanings are random and independent from the true state. Partisan leanings only affected the potential payoffs from the voting outcomes and had no impact on the belief elicitation task.

Model (2) of Table 1.3 shows that jurors no longer make this error after they have had a chance to deliberate. As would be expected of someone rationally responding to signals, the effect of the irrelevant partisan leaning on beliefs after deliberation is not statistically different from zero.

1.3.2 Votes

After participants have viewed the evidence and formed beliefs, the next step is to vote on the outcome.

Result 3. *Most jurors vote for the outcome they believe is more likely to be correct regardless of asymmetric incentives.*

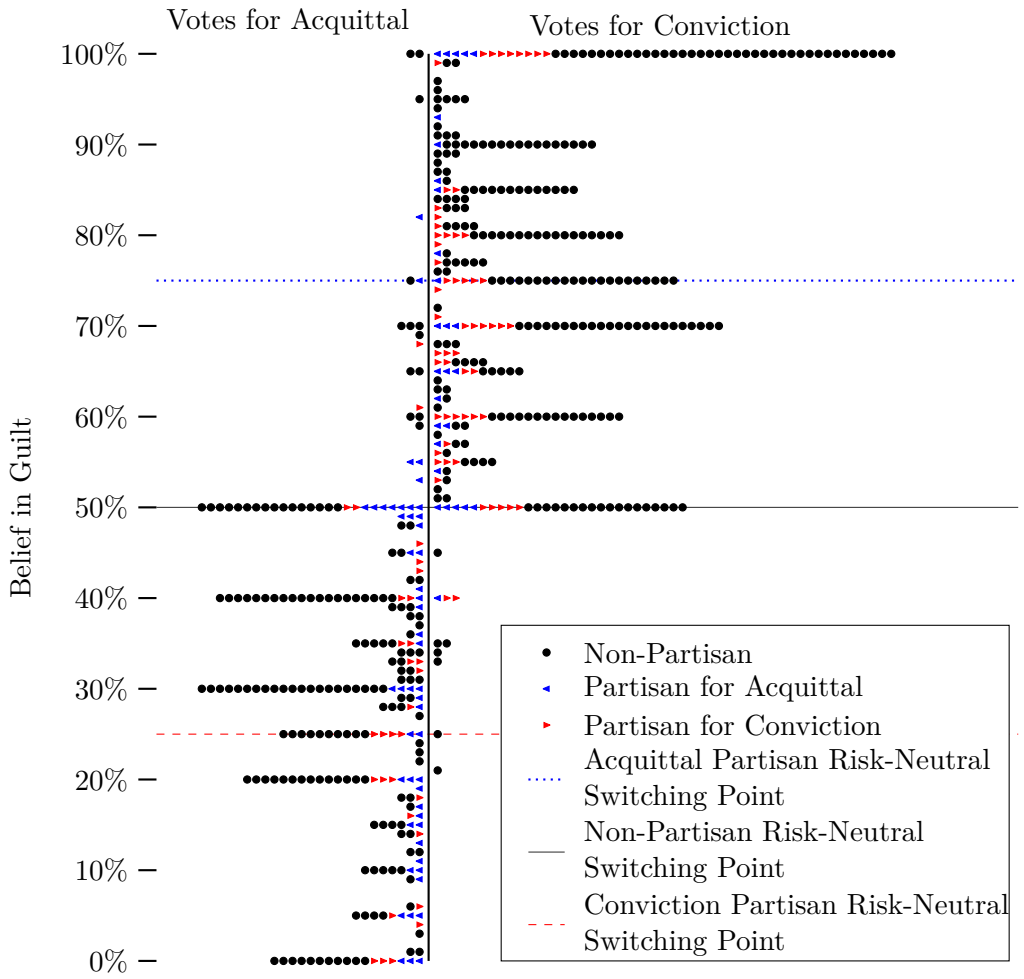
Before studying beliefs and belief updating, it is important to check whether beliefs correspond to voting behaviour. Most people vote for the outcome that they believe is more likely to be correct, even if the payoffs are asymmetric.

Figure 1.2 shows the relationship between beliefs (both initial and final) and votes. As the graph shows, the jurors tended to vote for acquittal when their belief in guilt was below 50% and for conviction when their belief was above 50%, regardless of partisan incentives. In treatments with symmetric incentives, only 20 of 408 (4.9%) votes were inconsistent with this pattern. Similarly, only 10 of 168 (6.0%) votes in the treatments with asymmetric incentives were inconsistent with this pattern, even though it would be rational for a risk-neutral juror to sometimes vote for the less likely outcome.

Hypothesis 3 does not hold because jurors are clearly not adopting the strategy that maximizes their expected payoffs. Jurors seem to maximize their chance of answering

[7] Contrast this with the final beliefs, where the mean belief of a guilt partisan was 55.2%, and the mean belief of an acquittal partisan was 50.4%. This difference is not statistically significant.

Figure 1.2: Votes by belief in guilt



Each point represents a belief/vote pair. Jurors' incentives imply risk-neutral switching points of 25%, 50%, or 75% depending on the partisanship they were assigned.

correctly regardless of the relative size of the payoffs. This would be consistent with either a very high degree of risk aversion (minimizing the chance of receiving the lowest payoff) or a non-monetary reward for being correct. However, either explanation undermines the reasoning in Section 1.1.1. If jurors only care about minimizing the odds of being wrong, partisan incentives are functionally identical to symmetric incentives, and truthful signal sharing is the optimal strategy in both cases.

1.3.3 Belief updating

After the first vote, the participants deliberate before voting and reporting their beliefs a second and final time.

Result 4. *When incentives are aligned, deliberation causes beliefs to update toward those of other jurors. This makes beliefs more accurate on average.*

Equation (1.3) established how a rational Bayesian would update their beliefs in response to independent signals. We cannot assume that jurors are Bayesian or that they interpret their signals as independent, and so I generalize Equation (1.3) by adding weight parameters, α and β, to each signal as well as a constant term, C (with $c \equiv \log C$).

The logit of some probability x is $\log[x/(1-x)]$. Juror i's belief in guilt is modeled as follows.

$$\frac{P(\text{Guilty}|s_i, s_j, s_k)}{P(\text{Innocent}|s_i, s_j, s_k)} = \left(\frac{s_i}{1-s_i}\right)^{\alpha} \left(\frac{s_j}{1-s_j}\frac{s_k}{1-s_k}\right)^{\beta} C \tag{1.7}$$

$$\text{logit}[P(\text{Guilty})] = \alpha \,\text{logit}(s_i) + \beta[\text{logit}(s_j) + \text{logit}(s_k)] + c \tag{1.8}$$

I refer to $\text{logit}[P(\text{Guilty})]$ as the juror's final belief, $\text{logit}(s_i)$ as juror i's own prior, and $\text{logit}(s_j)+\text{logit}(s_k)$ as others' priors. These each correspond to the reported beliefs expressed as logits. Finally, ε is an error term where $E(\varepsilon) = 0$.

$$\text{Final Belief} = \alpha\text{Initial Belief} + \beta\text{Others' Initial Beliefs} + c + \varepsilon \tag{1.9}$$

If the signals are independent and the jurors are Bayesian, then $\alpha = \beta = 1$ and $c = 0$. This corresponds to giving each signal full and equal weight ($\alpha = \beta = 1$) and not being biased in either direction ($c = 0$).

Model (1) of Table 1.4 estimates the parameters of Equation (1.9). Model (2) adds treatment dummies modifying both the constant and the weights shown in Equation 1.10.

$$\begin{aligned}\text{Final Belief} =& (\alpha_1\text{Subjective} + \alpha_2\text{Partisan} + \alpha_3\text{Framed})\alpha_0\text{Initial Belief} \\ &+ (\beta_1\text{Subjective} + \beta_2\text{Partisan} + \beta_3\text{Framed})\beta_0\text{Others' Initial Beliefs} \\ &+ c + \gamma_1\text{Subjective} + \gamma_2\text{Partisan} + \gamma_3\text{Framed} + \varepsilon\end{aligned} \tag{1.10}$$

Model (3) adds a term for the true state of the world (*Actual Guilt*) to detect when deliberation picks up accurate information not contained in the initial beliefs. This term takes a value of 1 when the defendant is guilty (or the true state is red) and 0 otherwise. This model is given in Equation 1.11.

$$\begin{aligned}\text{Final Belief} =&(\alpha_1\text{Subjective} + \alpha_2\text{Partisan} + \alpha_3\text{Framed})\alpha_0\text{Initial Belief} \\ &+ (\beta_1\text{Subjective} + \beta_2\text{Partisan} + \beta_3\text{Framed})\beta_0\text{Others' Initial Beliefs} \\ &+ (\delta_1\text{Subjective} + \delta_2\text{Partisan} + \delta_3\text{Framed})\delta_0\text{Actual Guilt} \\ &+ c + \gamma_1\text{Subjective} + \gamma_2\text{Partisan} + \gamma_3\text{Framed} + \varepsilon\end{aligned} \tag{1.11}$$

Table 1.4: Belief updating model

	Final Belief		
	(1)	(2)	(3)
Constant	0.227** (0.108)	−0.001 (0.190)	−0.312 (0.246)
Initial Belief	0.698*** (0.057)	0.674*** (0.118)	0.582*** (0.116)
Others' Initial Beliefs	0.243*** (0.037)	0.201** (0.086)	0.110 (0.087)
Subjective		0.626** (0.250)	0.694** (0.333)
Partisan		−0.089 (0.250)	−0.498 (0.320)
Framed		−0.052 (0.315)	0.022 (0.356)
Subjective×Initial Belief		0.141 (0.156)	0.237 (0.151)
Partisan×Initial Belief		−0.166 (0.147)	−0.159 (0.142)
Framed×Initial Belief		0.128 (0.152)	0.168 (0.149)
Subjective×Others' Initial Beliefs		0.113 (0.103)	0.209** (0.101)
Partisan×Others' Initial Beliefs		−0.075 (0.100)	−0.068 (0.098)
Framed×Others' Initial Beliefs		0.110 (0.107)	0.150 (0.108)
Actual Guilt			0.996** (0.406)
Subjective×Actual Guilt			−0.820 (0.515)
Partisan×Actual Guilt			1.191** (0.506)
Framed×Actual Guilt			0.048 (0.574)
Observations	288	288	288

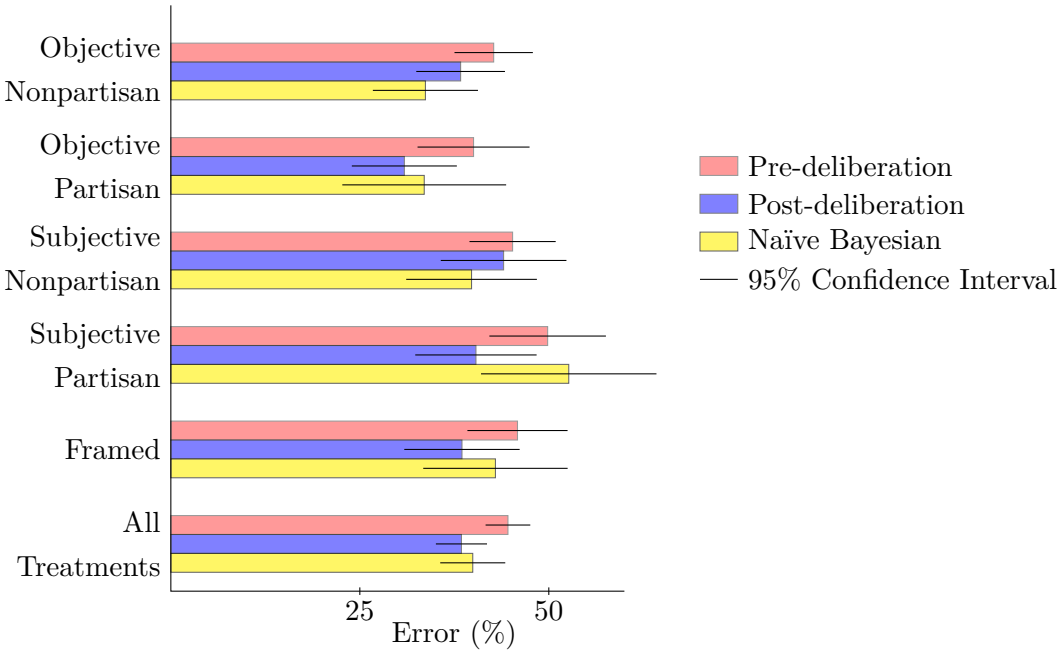
*p<0.1; **p<0.05; ***p<0.01

The positive and significant weight on the other jurors' priors shows that the jurors listened to each other and updated their beliefs to align with those of the group. While the

jurors were not fully Bayesian, e.g. putting more weight on their own priors than those of others, they still succeeded at incorporating information from other jurors.

Figure 1.3 shows how deliberation affects the accuracy of beliefs in all treatments, with two alternative prediction methods for comparison. The naïve Bayesian model follows Equation (1.3), simply combining all group members' prior beliefs as if they were independent signals of guilt. This is achieved by multiplying the odds ratios of all three jurors' prior beliefs together to find the final odds ratio.

Figure 1.3: Change in belief accuracy by treatment



Error is defined as the average distance between a belief or prediction and the outcome (1 for true, 0 for false).

Post-deliberation beliefs outperformed pre-deliberation beliefs in all treatments. This difference was significant in the objective partisan treatment ($p = 0.081$), the subjective partisan treatment ($p = 0.099$), and in all the treatments pooled together ($p = 0.008$). Furthermore, deliberation performs comparably to naïve Bayesian updating. Therefore, although deliberation does not look like Bayesian updating in Table 1.4, it still pushes beliefs in the correct direction on average. This is consistent with Hypothesis 4.

Result 5. *Jurors cooperate, share their signals, and update their beliefs as if their incentives were aligned, even when their incentives are not aligned.*

Model (2) in Table 1.4 shows no statistical difference between belief updating in the partisan and non-partisan treatments, falsifying the hypothesis that unaligned incentives hamper communication (Hypothesis 5). This makes sense in the context of Result 3. Theory

predicts that jurors will not truthfully share signals because they differ with respect to the threshold beliefs at which they switch from acquittal to conviction. However, Result 3 shows that most jurors have a threshold belief of 50% regardless of asymmetric incentives. Given that they share the same threshold, jurors have nothing to lose by truthfully sharing their signals.

This is consistent with the excessive truthfulness observed by Goeree and Yariv (2011) in their partisan treatments. In that experiment, jurors tended to truthfully share their signals even when theory predicted they would not.

Result 6. *Deliberation happens the same way regardless of the form of signals presented to jurors.*

Model (2) of Table 1.4 also shows that, with one exception, there are no significant differences in belief updating between the objective, subjective, and framed treatments.[8] The insignificant terms on the treatment dummies and their interactions show that treatment differences are not an important factor in belief updating. The regression results suggest that given three people with a set of initial beliefs, their deliberation will lead to a similar outcome whether they are deliberating over a murder case or an abstract question involving red and blue balls. The form of information does not seem to matter when it comes to belief updating under deliberation. This is consistent with Hypothesis 6.

Result 7. *Deliberation corrects the impact of irrelevant information.*

Two unexpected results stand out in the data: First, as shown in Figure 1.3, belief accuracy improved the most in partisan treatments. Second, Model (3) of Table 1.4 shows significant positive coefficients on *Actual Guilt* and the interaction term *Partisan*×*Actual Guilt*. This means that people updated their beliefs in the correct direction more than expected based on the beliefs in their group. This effect was stronger in partisan treatments, which featured the largest improvements in beliefs after deliberation.

Table 1.3 shows the results of two regressions: initial beliefs and final beliefs regressed on partisan incentives. The jurors were informed that their partisan leanings were randomly determined and independent of the true outcome. Nevertheless, their pre-deliberation beliefs were significantly swayed in the direction of their partisan leaning. This bias was corrected with deliberation; there was no significant impact of partisan leanings on post-deliberation beliefs.

This could be considered an instance of self-serving bias; partisan jurors stood to gain more in one scenario, and so they were biased toward seeing that scenario as more likely (see Brunnermeier and Parker, 2005). Alternatively, one could view it as a simple error.

[8]The one exception is a statistically significant bias in favour of updating toward the colour red in the subjective treatments. This result was unexpected, as I know of no theoretical reason to expect red and blue to be different.

Jurors were given a piece of information (their partisan bias) that some misinterpreted as a signal. In either case, their group members were able to correct their error, leading to an improvement in prediction accuracy.

1.4 Conclusion

This study makes two contributions to the study of deliberation. First, by adding belief elicitation to the experimental study of deliberation, the study adds insights into how people change their beliefs when they deliberate. Second, adapting the experiment to various settings shows that this type of experiment is robust concerning the form of information presented to jurors. Although the subjects' heuristic approach sometimes diverged from theory, it was at least consistent across different settings.

The results of the belief elicitation offer a potential answer to an apparent contradiction raised in the literature. Theoretical work by Austen-Smith and Feddersen (2006) shows that jurors with misaligned preferences have an incentive to lie. In contrast, this experiment finds no evidence of lying, consistent with past experiments (Goeree and Yariv, 2011; Le Quement and Marcin, 2020). A potential clue lies in the relationship between beliefs and votes. Jurors nearly always voted for the outcome they viewed as more likely, even if that outcome had a significantly lower upside payoff. This is consistent with the pattern Le Quement and Marcin (2020) found, wherein 80% of subjects adopted a heuristic of sharing their information and then voting for whichever outcome had the most evidence in its favour. It appears that part of the heuristic approach to deliberation is to always vote for the more likely outcome, not for the highest expected value. If all members of a jury behave this way, it violates the assumption of a *minimally diverse* committee, meaning a committee with a difference in decision thresholds. Austen-Smith and Feddersen (2006) proved that full truth-telling under a unanimity rule is not an equilibrium if and only if the committee is minimally diverse. Thus, the finding that the vast majority of jurors functionally have the same voting threshold is sufficient to explain why they tend to share their private signals truthfully.

This study's novel use of both a subjective task and a framed experiment for deliberation research has shown that neither subjectivity nor framing changes how people update their beliefs when deliberating.[9] This supports the external validity of the laboratory experiments in this area. All deliberation experiments rely on the assumption of similarity between the laboratory setting and the real-life settings where consequential deliberative decisions are made. In this case, the assumption holds.

This experiment unintentionally produced an opportunity for subjects to demonstrate motivated beliefs. Researchers have demonstrated that people form and hold motivated beliefs in a wide range of real-world and experimental settings (see Bénabou, 2015). In the

[9]With one exception: see footnote 8.

partisan treatments, subjects with a financial stake in one particular outcome believed that outcome was more likely even though the outcome was random and independent of their incentives (see Result 2). Intriguingly, this effect disappeared after the communication phase. Charness et al. (2018) demonstrates that there may be a strategic element to overconfident or motivated beliefs. People may self-deceive in order to be more convincing to others, correcting this after deliberation has already happened and there is no longer the possibility of deception. Alternatively, subjects may have simply conflated their partisan bias with a signal, correcting this error after discussing it with others. Further research could help to explore whether group discussion consistently corrects motivated and overconfident beliefs across a wider range of settings.

Chapter 2

Instructions

David J. Freeman, Erik O. Kimbrough, Garrett M. Petersen, and Hanh T. Tong[†]

Abstract

A survey of instruction delivery and reinforcement methods in recent laboratory experiments reveals a wide and inconsistently-reported variety of practices and limited research evaluating their effectiveness. Thus we experimentally compare how methods of delivering and reinforcing experiment instructions impact subjects' comprehension and retention of payoff-relevant information. We report a one-shot individual decision task in which non money-maximizing behaviour can be unambiguously identified and find that such behaviour is prevalent in our baseline treatment which uses plain, but relatively standard experimental instructions. We find combinations of reinforcement methods that can eliminate half of non money-maximizing behaviour, and we find that we can induce a similar reduction via enhancements to the content of instructions. Residual non money-maximizing behaviour suggests this may be an important source of noise in experimental studies.

[†]This paper was originally published in the Journal of the Economic Science Association. Address: Freeman, Petersen and Tong, Department of Economics, Simon Fraser University, 8888 University Dr, Burnaby, BC V5S 1A6, Canada. Kimbrough, Smith Institute for Political Economy and Philosophy, Chapman University, One University Drive, Orange, CA 92866, USA. We thank Jim Sylvester for programming the experiment, Cameron Young for research assistance, and Bob Slonim, two anonymous referees, Kevin Laughren, and seminar and conference audiences at Simon Fraser University, Chapman University, and the 2017 Economic Science Association World Meetings in San Diego for helpful comments. Freeman thanks Simon Fraser University for research funding (SSHRC-SFU Institution Grant SMALL-2013-631004). Kimbrough thanks SSHRC for additional funding (SSHRC Insight Grant 435-2015-0798). This study has been vetted by the Simon Fraser University Office of Research Ethics (study 2014s0066).

JEL classification: C91

Keywords: Attention, Comprehension, Instructions

2.1 Introduction

Experiments start by providing instructions designed to ensure that subjects understand how their actions and others' actions determine payoffs. Such understanding is crucial to the economic interpretation of subjects' behaviour – without it, the experimenter has lost control (Smith, 1982). Almost from the field's inception, experimental economists have recognized that the effectiveness of instructions in establishing understanding may depend on how they are delivered and reinforced (Fouraker and Siegel 1963). Prominent textbooks give detailed guidelines on how to deliver instructions and suggest complementary methods to increase subjects' comprehension, including reading instructions aloud and using demonstrations, quizzes, and practice rounds (Friedman and Sunder 1994, Davis and Holt 1993, Cassar and Friedman 2004). Casual observation suggests wide variation in how practitioners deliver instructions and use reinforcement methods. We review the methods for delivering and reinforcing instructions as reported in experimental studies recently published in six leading journals and confirm this observation. We find that almost all experimenters complement their instructions with at least one reinforcement method, though the methods used vary substantially. This suggests that experimental economics lacks clear norms for how instructions ought to be delivered and reinforced. Troublingly, we were unable to classify roughly 22% of papers because they failed to provide sufficient details on their methods.

Despite observed variation in practices, there is scant evidence comparing their effectiveness. Thus we conduct an experiment to evaluate the impact of methods of delivering instructions and reinforcing their content on behaviour. We study a one-shot timing decision in which each subject is performing a default Task 1 for money and must decide when (or whether) to switch over and complete Task 2. Task 2 can be performed at most once, and the subject is paid the most for doing it at the correct time and least for doing it earlier. Moreover, the subject is better off not doing Task 2 at all than doing it too early. This information is explicitly stated in the instructions. Doing the task too early – *non money-maximizing behaviour* (NMB) – could reflect idiosyncratic preferences, or result from a failure to comprehend or retain information from the instructions. Variation in NMB across treatments, which hold the distribution of preferences constant in expectation, thus reflects variation in comprehension and retention. For most treatments, we hold constant the content of instructions and vary how instructions are delivered and reinforced. We include one additional treatment with enhanced instructions as a robustness check.

In our first treatment subjects complete self-paced computerized instructions including practice rounds and then take a comprehension quiz before beginning the study (providing us an alternative measure of their comprehension upon completion of the instructions).

Nearly half of subjects in this treatment do the task too early, exhibiting NMB. A second treatment provides subjects with the quiz answers, and this generates a moderate, but statistically insignificant reduction in NMB. We thus study the additional impact of introducing monetary incentives for quiz performance, of going through the computerized instructions twice (both before and after the quiz), and of providing paper instructions alongside computerized instructions. We find that all three of these treatments lead to significant improvements relative to the baseline – but each only eliminates about half of the observed NMB, as does our treatment with enhanced instructions.

By studying an individual decision task, our experiment eliminates strategic and other-regarding motives that might confound the identification or interpretation of NMB. By studying a one-shot decision without feedback, we obtain a clean measure of understanding and retention of the instructions that is not confounded by learning. We are aware of two existing papers that have studied the impact of instruction delivery and reinforcement on play in repeated public goods games (Bigoni and Dragone, 2012; Ramalingam et al., 2018).[1] The more relevant of these is Bigoni and Dragone (2012), who find that shortened on-screen instructions led to lower quiz scores and longer response times as compared to their baseline paper instructions, shortened paper instructions, and shortened on-screen instructions with active examples requiring subject input. However, they find no effect of instructions on observed behaviour.

2.2 Literature Survey

We report how instructions are delivered and reinforced in 260 experimental studies published between January 2011 and December 2016 in Experimental Economics and five prominent general interest economics journals. We selected all papers in these journals that contained at least one lab experiment in which participants were given instructions on the experimental procedure. For each paper, we checked whether instructions were delivered on paper, on screen, both, or neither. We also recorded the use of various practices intended to reinforce the content of the instructions, including reading the instructions aloud, demonstrations, practice rounds, and pre-experiment quizzes. Since ensuring subjects' initial comprehension may be particularly important when experiments are one-shot or provide limited feedback, we further classified the nature of each experiment based on whether or not a main task was one-shot, and whether or not subjects received feedback. This allows us to assess whether experimenters adapt their instruction protocols to the nature of the

[1]Our discussion here is restricted to instruction delivery and reinforcement. We have little to say about how variation in the content of the instructions may affect behaviour, by providing or failing to provide subjects with payoff-relevant information, or alternatively by influencing the framing of the experimental task. See Alekseev et al. (2017) for a discussion of the use of context in instructions. See also Converse and Presser (1986) for a discussion of effective survey design which offers potentially useful guidance for economists.

task being studied. Details of our classification procedure are given in Appendix A. The results of our survey are given in Table 2.1.

We were unable to determine how instructions were delivered in 22% of the studies we reviewed. If behaviour is sensitive to how instructions are delivered, this oversight hampers replication. Of the remaining 204 studies, 61% deliver instructions exclusively on paper, 24% deliver instructions exclusively on screen, while another 5% use both. We find this noteworthy since the majority of these experiments are themselves computerized. The remaining 10% of these 204 studies use neither paper nor computer instructions. Most such studies are lab-in-the-field experiments studying non-student populations and deliver instructions orally along with some of the reinforcement methods discussed below. We suspect that experimental economists' revealed preference for paper instructions is driven by the fact that subjects can refer back to them throughout the experiment, which may not always be the case with computer instructions. This may mitigate subjects' tendency to forget important information.[2]

85% of all studies use at least one method of reinforcement which suggests that experimenters are almost universally concerned about subject comprehension and retention. Instructions are read aloud in 54% of studies. We find that 57% of studies use demonstrations or practice rounds to reinforce subject understanding of the experiment. Examples of such practices include physical demonstrations of how risk will be resolved,[3] guided examples of possible actions and their consequent outcomes, and unpaid practice rounds. Of the studies that use at least one of these forms of reinforcement, 80% use guided demonstrations or guided practice rounds, and 42% use unguided practice rounds; some studies use both.

In addition to reinforcing the content of instructions, experiments can also test subjects' comprehension thereof with pre-experiment quizzes (39% of studies). At least 63% of these reinforced understandings and corrected misunderstandings by providing answers to the quiz, and 41% required a perfect score to commence the experiment. Only three of the studies paid subjects for quiz performance. We note that 35% of studies that used a quiz did not clearly report whether or how subjects were given feedback on the quiz.

Given our prior that reinforcement may be especially important when feedback is limited, we find it surprising that one-shot experiments less frequently incorporate practice or demonstrations ($\rho = -.19$, $p < .01$, $n = 260$) and quizzes ($\rho = -.15$, $p = .02$, $n = 260$) in their instructions; see Appendix A for more detail.

Our survey reveals wide variation in how experimenters deliver and reinforce instructions. Nevertheless, there are commonalities which seem to reflect some notion

[2]Reading instructions aloud and/or publicly distributing paper instructions may also help establish common information in strategic settings (Friedman and Sunder 1994, p. 77).

[3]Davis and Holt (1993, p. 23) and Friedman and Sunder (1994 p. 67) suggest that the use of physical randomization devices may enhance credibility.

Table 2.1: Instruction delivery and reinforcement in economics experiments

Reinforcement	Delivery method: Computer only	Paper only	Computer and Paper	Neither	Unclear	Total
Total	**48**	**124**	**11**	**21**	**56**	**260**
Read aloud	19	79	4	21	17	140
Practice/Demonstration	30	63	10	15	29	147
Demo or guided practice	21	56	8	13	19	117
Unguided practice	16	22	4	4	16	62
Quiz	16	54	8	6	17	101
Feedback	10	35	5	4	10	64
Incentive	0	3	0	0	0	3
Require 100%	5	23	3	3	7	41
Feedback unclear	5	18	3	2	7	35
One-shot	15	43	4	12	10	84
Feedback between decisions	24	73	7	6	42	152

Each entry is the number of papers classified in that respective category. Indented categories are subsets of the preceding non-indented category.

of 'best practices.' Few studies have tested whether current practices are effective – our experiment is designed to fill this gap.

2.3 Experimental Design

2.3.1 Overview of Experiment

We design a one-shot, individual choice experiment in which each subject performs two tasks, a base task which provides a low flow of payoffs throughout the experiment, and a second task which can only be completed once and results in a potentially large lump-sum payoff. The amount of the lump sum depends on the time at which they initiate the second task. Doing the second task too early results in a lower payoff than doing it at the right time (or not doing it at all).

Task 1 is the Poodle Jump game (based on a popular mobile game Doodle Jump), where players guide a bouncing poodle up a series of platforms by pressing two buttons. When a subject misses a platform, the poodle falls to the ground and the game restarts with no penalty. Each participant receives \$0.25 per period of Task 1, so long as they jump a minimum cumulative height. This height was chosen so that it would be trivially easy to complete but not automatic – effectively guaranteeing an attentive subject this payment each period.[4]

Task 2 is a simplified version of the slider task (Gill and Prowse, 2012). Players can switch from Task 1 to Task 2 at any time by pressing the 'j' key, but they can only do this once. In the slider task, players are presented with four sliders which can be moved from zero to 100. The task is successfully completed when all four sliders are dragged to 50 and the player clicks "Continue."[5] Task 2's payoff depends on when the subject presses 'j'. For the first 21 periods, each period being one minute long, it pays \$0.20. However, in period 22 it jumps to \$7, falling to \$4 in period 23, then dropping by \$0.50 in every period thereafter until period 30 when the experiment ends. These payoffs are demonstrated in Figure 2.1. Doing Task 2 in period 22 maximizes a subject's payoff; whereas, doing it before period 22 minimizes a subject's payoff. If a subject fails to do Task 2 in period 22, they would always earn higher payoffs by doing it as soon as possible thereafter.

The challenge for subjects is to recognize and remember the correct time to press the 'j' key to complete Task 2, given the attention required to successfully complete Task 1 in each period. However, subjects have strong incentive to complete Task 2 at the right time: doing Task 2 at the right time raises payoffs by \$6.75 relative to not doing it at all, and by a minimum of \$3 compared to completing it at any other time. Moreover, doing it

[4]Only 5 out of 308 subjects ever failed to attain the required height in a period; 4 did so once and one subject did so twice. These failures account for only 0.1% of all Poodle Jump periods.

[5]Only one subject started but failed to complete the slider task.

Figure 2.1: Screenshot showing how payoffs were described to subjects

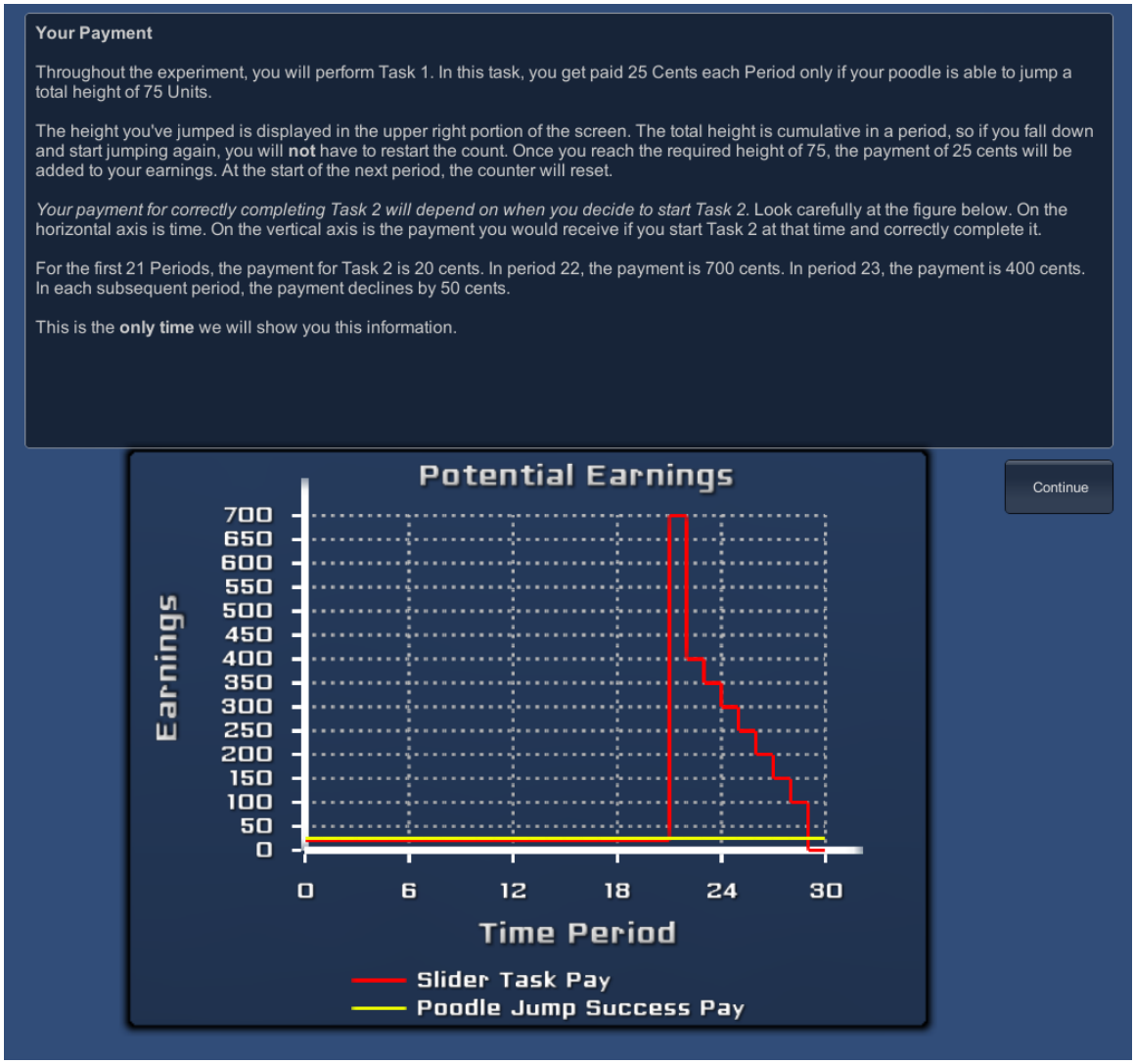


Table 2.2: Summary of treatments

Treatment	Quiz	Answers	Additional Reinforcement	# of Subjects
NO QUIZ	No	No	No	43
QUIZ	Yes	No	No	76
ANSWERS	Yes	Yes	No	36
INCENTIVE	Yes	Yes	Pay 0.50 CAD per correct quiz answer	38
TWICE	Yes	Yes	Instructions restarted unexpectedly	38
PAPER	Yes	Yes	Instructions duplicated in paper printout	40
ENHANCED	Yes	No	Only through enhanced on-screen instructions	37

before period 22 leads the subject to forgo the opportunity to do it at the optimal period or thereafter, and also results in a lower payoff than never doing Task 2. Thus, doing Task 2 before period 22 precludes the subject from maximizing their monetary payoffs. We use the NMB acronym to refer to such behaviour below.

NMB can thus reveal that a subject failed to comprehend or retain a particularly key piece of payoff-relevant information from the instructions.[6] As hinted at earlier, our design restricts the set of possible preference-based explanations for NMB. Moreover, since we sample subjects from the same distribution of preferences in each treatment, variation in NMB across treatments identifies changes in comprehension and retention.

2.3.2 Treatment Design

We employ a between-subjects design with seven treatments. We study the effectiveness of different ways of delivering and reinforcing the experiment's instructions on NMB using our aforementioned measure. Many experimenters implicitly assume that subjects fully understand their instructions. If this is true, we should not observe any difference between treatments. However, if subjects do not always comprehend or retain information from the instructions there is the potential for variation in delivery and additional reinforcement to reduce NMB. Our treatments test the impact of various more-or-less standard procedures employed by experimenters to improve comprehension and retention. All treatments are summarized in Table 2.2. All treatments started with a common set of self-paced on-screen instructions, which included a graphical explanation of payoffs as well as reinforcement from practice rounds for both tasks and practice switching between tasks.

The NO QUIZ treatment presents the instructions on screen with no additional reinforcement. The NO QUIZ treatment gives us information on NMB when subjects read instructions on their own.

[6]We note that neither a subject who understood and retained this information but simply forgot to switch nor a subject who (for whatever reason) did not understand this information but only switched at or after period 22 would be coded as exhibiting NMB by this measure.

The QUIZ treatment was identical to the NO QUIZ except that each subject completed a six question comprehension quiz on paper at the end of the on-screen instructions; subjects were informed that there would be a quiz prior to beginning the instructions, but no feedback was given on the quiz. The QUIZ treatment allows us to assess whether the presence of the quiz affects NMB, and the quiz itself gives a secondary measure of comprehension. When we analyze our data, we use this as our baseline treatment for comparison to the other treatments below.

The ANSWERS treatment was identical to the QUIZ treatment, except that subjects were presented the answers to the quiz orally after all had completed it. This corrected possible misunderstandings revealed in quiz answers and reinforced key pieces of information from the instructions. As noted by Cassar and Friedman (2004), a quiz is a good way to "make sure that the subjects understand the rules" (p. 71); thus we expect providing the answers to the quiz will correct failures of comprehension or retention and reduce NMB.

The TWICE treatment was identical to the ANSWERS treatment except that after completing the quiz and answers, the experimenter unexpectedly restarted the instructions for the participants to work through a second time. This allowed subjects to further review any content they missed on the first go and provided additional reinforcement. As noted by Friedman and Sunder (1994), "[when] a subject does not seem to understand the instructions [...] the experimenter may reread the relevant part of the instructions or go through an example" (p. 77). Repeating the instructions TWICE achieves both of these objectives and thus should reduce NMB.

The INCENTIVE treatment was identical to the ANSWERS treatment except that subjects were paid \$0.50 for each correct quiz answer, and were informed of this before starting the instructions. We hypothesized that this would lead subjects to pay more attention to the material in the instructions, and make any mistakes from the quiz more salient, thereby improving understanding. Pay for performance is standard in experimental economics because economists believe it motivates subjects to think carefully and participate actively in experiments (Hertwig and Ortmann, 2001). By paying for performance on the quiz, we anticipate that subjects will exert more effort in carefully reading the instructions, thereby reducing NMB.

The PAPER treatment was identical to the ANSWERS treatment except that the experimenter also distributed paper printouts of the instructions (in addition to the on-screen instructions), which participants could keep and reference at any time, even while completing the quiz.[7] We thus expect PAPER to improve comprehension as measured by quiz scores and reduce NMB both for this reason, and through improving retention given the quiz score since written instructions are available throughout the session.

[7]The PAPER treatment potentially reduces forgetfulness since all relevant information is accessible throughout the experiment.

The ENHANCED treatment was identical to the QUIZ treatment but with enhanced on-screen instructions.[8] Compared to the other treatments, the on-screen instructions were lengthened from five to seven screens in length. In these enhanced instructions, Figure 2.1 appeared four times (instead of only once), and subjects were presented with four worked-out examples that explained the payoff that would result from different possible switching times. Unlike in our other treatments, the last page of the enhanced instructions included Figure 2.1, and each subject waited on that page while other subjects completed the instructions and while they completed the quiz. With the benefit of hindsight, we emphasized the details we knew past subjects had failed to grasp. This treatment is also consistent with the advice of Friedman and Sunder (1994), applied between-subjects, and we expect the ENHANCED instructions to similarly reduce NMB.

For reasons explained above, we hypothesize that each additional form of reinforcement reduces NMB. Specifically, we conjectured that having a QUIZ would have a similar level of NMB as NO QUIZ, but relative to these treatments, ANSWERS would reduce NMB, each of our remaining interventions on top of that (INCENTIVE, TWICE, and PAPER) would further reduce NMB, and ENHANCED would also reduce NMB relative to QUIZ. We hypothesized that higher quiz scores will be associated with lower rates of NMB, and that in the INCENTIVE, PAPER, and ENHANCED treatments most or all reductions in NMB are reflected in higher quiz scores, while the ANSWERS and TWICE treatments reduce NMB given quiz scores.

Our experiment differs from existing studies on instructions in two regards. First, this is an individual decision task, so there is neither complexity from strategic behaviour nor other-regarding concerns. Second, it is a one-shot task – each subject can only press 'j' once – so participants who fail to understand the instructions cannot learn through trial and error. These features allow us to cleanly identify NMB and attribute variation in NMB to variation in the delivery and reinforcement of instructions. Nonetheless, we believe that our experiment provides a good analogy to other experiments, particularly those where a decision of interest is only one of multiple decisions the subject makes. We also conjecture that more complicated experiments face at least as much risk of misunderstanding as exists in our simple experiment (even if most existing experiments are unable to diagnose it).

2.3.3 Procedures

Upon entering the lab, the experimenter assigned participants to visually isolated computer terminals. Participants were told not to interact with one another for the duration of the experiment. In all treatments, participants were informed that they would be given a set of instructions followed by an experiment in which they could potentially earn a significant amount of money; in the treatments with a quiz, they were also informed that there would be

[8]The ENHANCED treatment was added later on a suggestion from the editor.

a quiz at the end of the instructions; subjects in the INCENTIVE treatment were informed that they would be paid for their quiz performance above and beyond their earnings from the experiment. The experimenter then started the self-paced on-screen instructions which included a written description of the tasks and the payoff structure, practice rounds of both tasks, practice switching between tasks, and a graphical illustration of the payoffs to both tasks in each period (a full copy of the instructions are presented in Appendix B). Once all participants completed the instructions, the experimenter distributed the quiz in the QUIZ, ANSWERS, INCENTIVE, TWICE, PAPER, and ENHANCED treatments; the correct answers were revealed after all participants had completed the quiz except in the QUIZ and ENHANCED treatments. In the TWICE treatment, subjects completed the on-screen instructions a second time, including practice rounds. Then the experiment started. At the end of some sessions, we conducted a post-experiment questionnaire (Appendix D).[9]

We recruited 308 participants to 45 sessions through Simon Fraser University's CRABE recruiting system, with no subject participating in more than one session. Each session lasted under an hour. Average earnings were 18.37 CAD including a 7 CAD show-up payment. We collected no other demographic data nor other behavioural measures.

2.4 Results

We use a subject's decision to do Task 2 at any time before period 22 as NMB, which is our behavioural measure of their failure to pay attention to, comprehend, absorb, or retain information from the instructions. Table 2.3 shows the share of NMB by treatment. All p-values reported below are two-sided.

Finding 1: NMB is prevalent.

In our NO QUIZ and QUIZ treatments, 44% and 47% of subjects exhibited NMB by doing Task 2 before period 22. This is despite the fact that these treatments include both demonstrations and practice periods. Even in our most effective treatment, the corresponding share is 18%. These findings suggest that failures to comprehend or retain information from instructions may be an important source of noise.[10] This justifies concern about the effectiveness of instruction delivery and reinforcement methods.

Finding 2: Combining reinforcement methods reduces *NMB*.

We find that additional reinforcement reduces NMB: we reject the joint hypothesis that NMB occurs at the same rate across all treatments (Fisher's exact test, $p < .01$, $n = 308$).

[9]We have responses from 72 subjects because this was added at the suggestion of a referee.

[10]In Appendix C, we show that we find similar results if we account for trembles by defining NMB based on doing Task 2 before period 21.

Table 2.3: *Non Money-maximizing behaviour* across treatments

	NO QUIZ	QUIZ	ANSWERS	INCENTIVE	TWICE	PAPER	ENHANCED
NMB	.442	.474	.333	.237	.184	.225	.216
Quiz Score (avg.)	n/a	4.10	4.06	4.32	4.53	5.43	4.59
QUIZ	.849						
ANSWERS	.362	.220					
INCENTIVE	.064	.016	.442				
TWICE	.017	.004	.186	.779			
PAPER	.062	.010	.316	1.00	.781		
ENHANCED	.056	.013	.302	1.00	.779	1.00	

First row reports the fraction of NMB by treatment.
Second row reports the average quiz score by treatment. Remaining entries report a
p-value from a Fisher's exact test of differences in NMB between treatments.

Compared to NO QUIZ and QUIZ, we observe somewhat less NMB in the ANSWERS treatment (33%), but we do not detect any statistically significant differences between these treatments (Fisher's exact test of equal NMB rates across these treatments, $p = .35$, n =155). In each of the INCENTIVE (24%), TWICE (18%), and PAPER (23%) treatments that provide additional reinforcement, subjects exhibited significantly less NMB than in the QUIZ treatment (Fisher's exact tests, $p < .02, .01, .01$, $n = 114, 114, 116$ respectively). While the ENHANCED treatment (22%) reduces NMB (Fisher's exact test, $p = .01$, $n = 113$), it does not eliminate it.[11] Our findings suggest that more detailed instructions and extensive reinforcement each improve comprehension and retention of the instructions.

Finding 3: Lower quiz scores are associated with NMB. Providing quiz answers while also making incorrect answers salient can reduce NMB among lower performers.

Quiz scores provide an alternative measure of subject comprehension immediately after the instructions. In the QUIZ treatment which provides neither feedback nor additional reinforcement, quiz score and NMB are negatively related (Goodman-Kruskal γ, $p < 0.01$, $n = 76$); indeed 13 of 76 subjects had a perfect score on the quiz, and none of them subsequently exhibited NMB in the experiment. In fact, across all of our treatments we find it striking that only one of the 73 people with a perfect quiz score exhibited NMB.[12] This indicates that full *comprehension* at the completion of the instructions appears to be a sufficient condition for avoiding NMB in our experiment and that *retention* is a second-order issue.

Our quiz score data enable us to test whether the INCENTIVE, PAPER, and ENHANCED treatments improved subjects' comprehension as demonstrated on the quiz, compared to the pooled distribution of quiz scores from the QUIZ, ANSWERS, and TWICE treatments, which followed identical procedures up to the collection of the quiz.[13] Average quiz scores by treatment are reported in Table 2.3. To our surprise, neither the INCENTIVE nor the ENHANCED treatment significantly improved quiz scores (rank-sum tests, $p = .59, .14$, $n = 188, 187$, respectively). The PAPER treatment, which made the answers accessible to subjects during the quiz, improved scores significantly (rank-sum test,

[11] We cannot reject the hypothesis INCENTIVE, TWICE, PAPER, and ENHANCED lead to similar improvements (Fisher's exact test of no association, $p = .96$, $n = 153$).

[12] One person with a perfect quiz score in the TWICE treatment switched 28 seconds too early.

[13] We find no significant differences in the distribution of quiz scores in the QUIZ, ANSWER, and TWICE treatments (Kruskal-Wallis test, $p = .15$, $n = 150$).

Table 2.4: Treatment effects on Non Money-maximizing behaviour and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB		Quiz Score	NMB	
	(1)	(2)	(3)	(4)	
NO QUIZ	-0.128				n
	(-0.889, 0.632)				
ANSWERS	-0.588	0.051	-0.050	-0.138	
	(-1.424, 0.248)	(-2.884, 2.987)	(-0.586, 0.487)	(-0.308, 0.048)	112
ANSWERS × Quiz Score		-0.206		0.00715	
		(-0.941, 0.528)		(-0.069, 0.085)	
INCENTIVE	-1.065**	-2.531*	0.211	-0.219	
	(-1.948, -0.182)	(-5.332, 0.271)	(-0.354, 0.775)	(-0.392, -0.030)	114
INCENTIVE × Quiz Score		0.361		-0.028	
		(-0.257, 0.978)		(-0.111, 0.049)	
TWICE	-1.383***	-2.181	0.421	-0.255***	
	(-2.329, -0.436)	(-5.235, 0.873)	(-0.156, 0.998)	(-0.425, -0.065)	114
TWICE × Quiz Score		0.207		-0.057	
		(-0.467, 0.880)		(-0.145, 0.021)	
PAPER	-1.131**	7.334*	1.320***	0.134	
	(-2.009, -0.253)	(-0.810, 15.478)	(0.922, 1.718)	(-0.189, 0.343)	116
PAPER × Quiz Score		-1.485*		-0.177***	
		(-2.970, 0.001)		(-0.273, -0.085)	
ENHANCED	-1.182**	-0.557	0.489*	-0.188*	
	(-2.096, -0.269)	(-4.596, 3.482)	(-0.030, 1.008)	(-0.382, 0.013)	113
ENHANCED × Quiz Score		-0.116		-0.067*	
		(-0.993, 0.760)		(-0.151, 0.004)	
Quiz Score		-0.679***			
		(-1.053, -0.306)			
Intercept	-0.105	2.683***	4.105***		
	(-0.561, 0.350)	(0.993, 4.374)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

$p < 0.01$, $n = 190$), and the linear regression in Table 2.4, column 3 shows that PAPER had the largest effect on quiz score of all of our treatments.[14]

Quiz score data also allow us to further assess *how* our treatments reduce NMB. Goodman-Kruskal γ tests revealed that quiz score had a significant ($p < .05$ in each test, $n = 76, 36, 38, 40, 37$ respectively for each of QUIZ, ANSWERS, TWICE, PAPER, and ENHANCED) negative relationship with NMB in each treatment except INCENTIVE (where $p = .054$, $n = 38$) and NO QUIZ (where scores were not available). To decompose the extent to which treatment effects operate via (i.e. are *mediated* through) improved comprehension demonstrated on the quiz, we perform mediation analysis (applying the approach of Imai et al. 2010) in column 4 of Table 2.4, based on a model of NMB as a logistic-linear function of quiz score, treatment, and their interactions (column 2), and a linear regression to model treatment effects on quiz scores (column 3). The INCENTIVE and TWICE treatments have sizable and significant direct effects but insignificant and small mediated effects.[15] This indicates that these treatments primarily reduce NMB by clearing up (TWICE) and making salient (INCENTIVE) failures of comprehension demonstrated on the quiz. In contrast, the PAPER treatment has the largest mediated effect of all treatments, which is statistically significant, but only a small and insignificant direct effect beyond that. Mediated and direct effects of the ENHANCED treatment are each borderline insignificant, indicating a mix of both types of effects, but point estimates indicate a larger direct effect.

Robustness Checks Figure 2.2 shows empirical CDFs of completion times for Task 2, by treatment. For robustness, we show in Appendix C that we would arrive at similar qualitative conclusions to those reported in Table 2.4 using any of three alternative measures of NMB which vary the strictness of the criteria by which we classify behaviour as NMB.

Our post-experiment questionnaire was only partially able to diagnose causes of NMB in our experiment (see Appendix D for a full analysis). While subjects' responses are correlated with behaviour and quiz scores, they fail to provide any indication of the differences between the QUIZ and ENHANCED treatments in NMB revealed in the experiment.

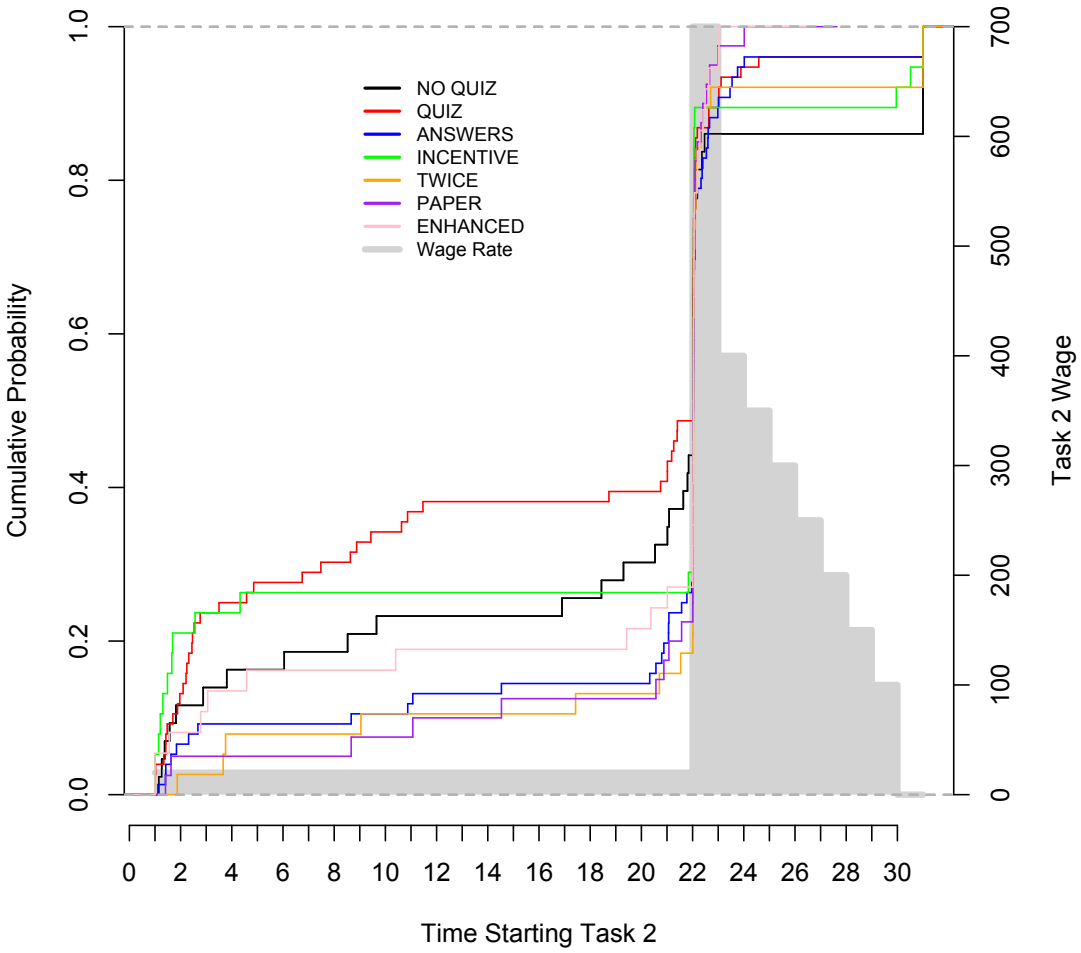
2.5 Discussion

Our experiments indicate that even when using combinations of reinforcement methods including demonstrations, practice periods, and a quiz, many subjects' behaviour reveals that they fail to pay attention to, understand, or retain information from the instructions. Combining these with further reinforcement methods reduced NMB, as did increasing the

[14]The positive effect of paper instructions on quiz performance is consistent with the evidence reported in Bigoni and Dragone (2012).

[15]In the case of TWICE, this is reassuring since any mediated effect can only arise due to sampling variation.

Figure 2.2: Empirical CDFs of Task 2 completion times, by treatment.



level of detail in the instructions. Each of these methods leads to a similar improvement but does not eliminate NMB.

In our setting, we feel confident attributing variation in the anomalous behaviour that we observe to a variation in the failure to understand or absorb the instructions. In other experiments designed to test for anomalous behaviour, the distinction between truly anomalous behaviour of interest and a failure to understand the instructions may not be so clearcut. This justifies a concern with how instructions are given and the use of behavioural checks of understanding. Our findings broadly suggest that experimenters' attempts to reinforce the instructions or make them more salient can be effective at reducing NMB. Note that though we are able to reduce NMB in our design, some residual NMB persists even in the best case. While the extent of such NMB is likely to vary with experimental context (e.g. subject pool, design, feedback), its presence is noteworthy and has implications for the power and interpretation of experimental tests.

Finally, our findings motivate advice on how to report and deliver instructions. First, experimenters should be aware that the way instructions are delivered and reinforced has consequences for behaviour. Second, we suggest providing paper instructions when possible, since this requires no extra lab time, is almost free, and is about as effective in reducing NMB in our experiment as other reinforcement methods. Third, we suggest that all experimental papers should clearly report how they deliver and reinforce instructions, as this can be crucial for close replication and interpretation.[16] Journals' efforts to require experimenters to share copies of their instructions are laudable, and these could be complemented by standardized reporting of how instructions are delivered and reinforced.

[16]For example, recent work by Chen et al. (2018) demonstrates, via new experiments following different instructions protocols, that a recent failed replication attempt arose because of differences in how instructions were delivered.

Chapter 3

Mobility as a Service Apps and Multimodal Transportation: Evidence From a Multimodal App

Hao Li, Garrett M. Petersen, and Fei Yu

Abstract

Recent years have featured rapid innovation in digital platforms connecting users with transportation services. Aggregator apps and Mobility as a Service (MaaS) platforms offer users the convenience of accessing multiple different services through a single app. These apps offer end-to-end route planning combining and comparing multiple different modes of transport. Some transit authorities have bet on these apps as a means to promote transit usage by encouraging ride hailing as an extension to fixed-route, fixed-schedule transit lines. Using data from a popular aggregator app, we test whether a multimodal route-planning service caused users to use combined routes featuring both ride hailing and transit. We find that ride-hailing trips connected with rail stops increased from 3.0% of trips to 5.5% among existing users. In areas where the feature supported bus connections, trips connecting to bus stops increased from 4.6% to 8.7% among existing users. Our results indicate that aggregators increase the degree of complementarity between ride hailing and transit.

The 21st century has seen rapid innovation in digital platforms for transportation. Starting in 2005 with the development of the General Transit Specification Feed (GTFS), transit agencies around the world have released open source information on their routes and vehicles. This has allowed mapping applications such as Google Maps to help users plan transit trips with accurate, up-to-date information. With the development of modern smartphones, beginning with the first iPhone in 2007, users could have transit information anywhere, any time through one of many mapping applications connected to GTFS feeds.

Once smartphones became ubiquitous, new platforms emerged to allow users to summon a vehicle through their phones. Ride hailing platforms Uber and Lyft both launched in San Francisco in 2010, and have emerged as major competitors to the taxi industry by connecting drivers with passengers through digital platforms. Others followed, with entrants to the space creating digital platforms for users to locate and rent other vehicles such as bicycles and scooters, as well as vans and larger vehicles offering shared service (i.e. "microtransit").

With a wide array of mobility services scattered across different platforms, a natural development was the creation of aggregator apps that could combine multiple forms of mobility operated by different organizations. With an aggregator app, users can easily compare all available modes of travel and pick the combination of speed, convenience, and cost that they prefer. If no one mode of travel is best, the app can help users plan a multimodal route involving multiple services. This is particularly useful for combining the efficiency and low cost of mass transit with the flexibility of a car by linking transit and ride hailing trips together. These aggregators are the most basic form of Mobility as a Service (MaaS), digital platforms combining multiple transportation services with the goal of competing with private cars. Transit agencies have taken notice of this development, and some are betting big on MaaS as a means to promote transit by extending the range of fixed-route, fixed-schedule transit lines (Goodall et al., 2017). To compete with the flexibility of private cars, MaaS needs to offer a convenient means of getting anywhere a car can go. Multimodal trips fill a niche by facilitating trips that are not well-served by transit or ride hailing alone, making them an important part of the MaaS ecosystem.

We study whether the integration of different modes of transport within an aggregator app can promote the use of ride hailing as a first- and last-mile solution for transit. A popular mapping app added a multimodal route planning feature in November 2018, with bus connections supported in select cities and rail connections supported everywhere. We observe the change in multimodal behaviour by measuring the change in the proportion of ride hailing trips starting or ending at transit stops. We find that this feature significantly increased multimodal trips. Among existing users of the app, ride-hailing trips connected with rail stops increased from 3.0% of trips to 5.5%. In areas where the feature supported bus connections, trips connecting to bus stops increased from 4.6% to 8.7% among existing users. Meanwhile, cities where bus travel was not supported by the multimodal feature saw no change in the rate of bus connections.

Our results provide evidence that the route-planning element of MaaS succeeds in promoting multimodal travel. We model the commuters' decision process to show how a reduction in the cost of planning corresponds to a reduction in the overall cost of travelling via transit and ride-hailing. By reducing the barriers to multimodal travel, the aggregator implicitly makes transit and ride hailing more appealing relative to private car travel.

3.1 Background and Lit Review

MaaS not only aggregates many different modes of transport into a single digital platform but also provides novel payment models (Sochor et al., 2015; Jittrapirom et al., 2017). Sochor et al. (2018) classifies MaaS into five levels from 0 to 4. The levels are cumulative. Level 0 is simply the absence of integration between modes, level 1 is integration of information, level 2 is integration of booking and payment, level 3 is integration of service as a comprehensive alternative to private car ownership, and level 4 is integration of societal goals through coordination with public authorities. The most developed existing example of MaaS is Helsinki's Whim app, which allows users to book and pay for many different modes of transport across the city through a single platform (Goodall et al., 2017). MaaS advocates hope to create more a more sustainable transportation system with less reliance on cars. To that end, Anagnostopoulou et al. (2020) use strategic messages within an aggregator app to nudge users towards more sustainable modes of transport.

Aggregators increase the convenience of multimodal trips by offering end-to-end route planning. Without this feature, a user could manually plan their bus or train route from a stop near their origin to a stop near their destination. Then they could summon a ride-hailing vehicle to connect the various legs of the journey. To do this users are required to navigate back and forth between a GTFS-connected mapping app like Google Maps and a ride-hailing app like Uber. Sometimes, they may even use more than one ride-hailing app to compare the price and the time efficiency of ride hailing. This inconvenient process did not allow for easy planning of multimodal routes, nor did it allow for easy comparison of multimodal routes against other options. Nonetheless, many users endured this technical inconvenience to connect ride hailing with transit absent an aggregator app. Hall et al. (2018) find that Uber is a complement to public transit on average, but more so in larger cities. There is mounting evidence that ride hailing causes fewer people to own cars, as many substitute towards a combination of ride hailing and transit to satisfy their urban mobility needs (Clewlow and Mishra, 2017; Hampshire et al., 2017; Ward et al., 2019).

Ride hailing has been controversial since its introduction, with many cities initially fighting to stop its entry (Spicer et al., 2019). The industry has clear benefits to consumers. Cohen et al. (2016) estimate that UberX generated $6.8 billion in consumer surplus in the United States in 2015. However, critics worry about negative secondary effects. Erhardt et al. (2019) found that Uber and Lyft greatly increased congestion in San Francisco. They estimate that ride hailing caused the average speed of vehicles driving in San Francisco to fall by 2 miles per hour (a 9% decline) due to increased congestion. Ride hailing has had an ambiguous impact on traffic fatalities, with some research finding increased fatalities associated with ride hailing companies entering a market (Barrios et al., 2019), some finding decreased fatalities (Huang et al., 2019), and others finding no change (Brazil and Kirk, 2016; Kirk et al., 2020).

Although ride hailing remains controversial, the debate on whether to allow its operation has been largely decided. With Uber operating in over 900 cities worldwide, ride hailing is here to stay. The relevant policy questions today revolve around how to integrate ride hailing and other digital platforms for transportation into cities in a way that best serves those cities' residents. This means finding policies that can enhance the social benefits of these platforms while reducing their social costs. Our results show that aggregators integrating different modes of transport such as ride hailing and transit promote multimodal trips. Basso and Silva (2014) show that the marginal social cost of transit is below the marginal social benefit given current policies, so promoting transit through subsidies or other policies can produce net social benefits. Adopting policies that promote aggregators and MaaS can help increase transit ridership by making ride hailing and transit more complementary.

3.2 Theory

A commuter wants to travel from a set origin to a set destination while minimizing the cost, $c(\cdot)$, of the trip. Cost here includes both monetary and time costs. Our data only includes ride hailing trips that may or may not connect to transit stops. Therefore, we focus on the case where there are two possible modes of transport: ride hailing (r) or multimodal (m). The commuter does not know their costs initially, but he has the option to look up a route for either or both modes in an app, at which point he will know that route's cost. The commuter cannot take a mode without first looking up the route. Checking the app takes time, which the commuter values at a constant value, s.

Assume that the distribution of trip costs for both modes follow a Poisson distribution with known parameters of λ_r and λ_m.[1] Apps function by finding the lowest-cost option and serving that to the user. Whereas basic apps require the user to search each mode separately, incurring a search cost of s each time, aggregator apps search all modes simultaneously incurring the search cost only once. Equation (3.1) gives the probability that the lowest available cost, c_i is less than or equal to p given some value of λ_i.

$$P(\mathrm{c}_i \leq p) = 1 - e^{-\lambda_i p} \tag{3.1}$$

Solution with an aggregator The aggregator makes the commuter's problem trivially solvable. The commuter simultaneously checks both values of c_i for a cost of s, then chooses the lower-cost option. His total cost is given by cost $= s + \min\{c_r, c_m\}$, which follows a cdf of

[1]We do not have data that would allow us to estimate the distribution of costs for any given mode, so we make this parametric assumption to make the model tractable.

$$P(\text{cost} \leq p) = \begin{cases} 0 & \text{if } p < s \\ 1 - e^{-(\lambda_r+\lambda_m)(p+s)} & \text{if } p \geq s \end{cases}. \tag{3.2}$$

Therefore the expected cost is

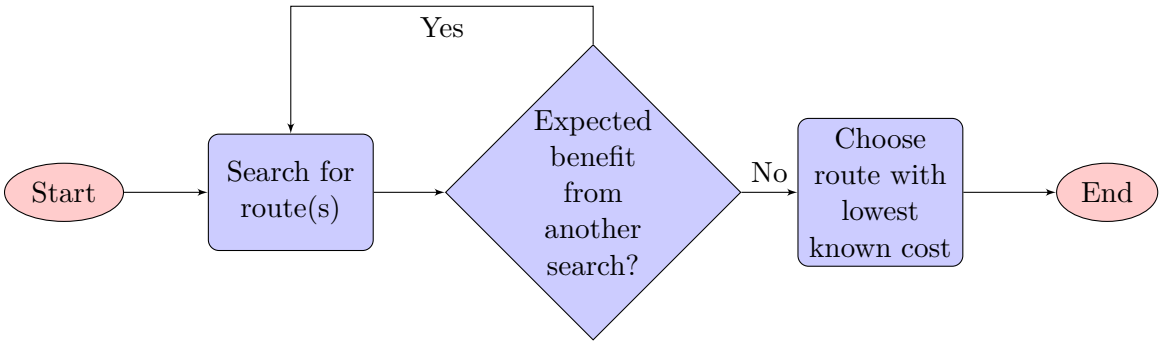
$$E(\text{cost with aggregator}) = s + \int_0^\infty p(\lambda_r + \lambda_m)e^{-(\lambda_r+\lambda_m)p}dp = s + \frac{1}{(\lambda_r + \lambda_m)}. \tag{3.3}$$

The probability of choosing a multimodal trip over ride hailing is given by

$$P(c_m \leq c_r|\lambda_r, \lambda_m) = \int_0^\infty (\lambda_r e^{-\lambda_r p})(1 - e^{-\lambda_m p})dp = \frac{\lambda_m}{\lambda_m + \lambda_r}. \tag{3.4}$$

Solution without an aggregator When the commuter does not have access to an aggregator, his problem becomes more difficult. He must choose which mode to investigate first, then, conditional on his observation of c_i for that mode, decide whether to observe the other mode. Figure 3.1 shows the commuter's decision process.

Figure 3.1: Commuter's decision process



The solution to the commuter's problem can be broken down into two answers: Which mode should be observed first, and under what condition should the other mode be observed.

Assume that mode i has been observed, and that mode j is unknown. If the commuter does not observe mode j, his cost is $s + c_i$. If the commuter does observe mode j, his cost is $2s + \min\{c_i, c_j\}$. The expected value of observing mode j can be found by taking the difference between the two costs, and integrating over the probability distribution of c_j.

$$EV(\text{observe } j|c_i, \lambda_j) = -s + \int_0^{c_i} (c_i - p)\lambda_j e^{-\lambda_j p}dp = -s + c_i + \frac{e^{-c_i\lambda_j} - 1}{\lambda_j} \tag{3.5}$$

Equation (3.5) implies that the expected value of observing mode j is increasing in both λ_j and c_i. The commuter will observe mode j if and only if this expected value in Equation

(3.5) is at least zero. Expressing this as a value of c_i, we get $EV(\text{observe } j|c_i, \lambda_j) = 0$ when

$$c_i = \bar{c} = \frac{W\left(-e^{-\lambda_j s-1}\right) + \lambda_j s + 1}{\lambda_j} \tag{3.6}$$

where $W()$ is the Lambert W function.[2] To fully characterize the solution, we need to determine which mode the commuter searches first. The commuter minimizes expected cost by first observing the mode with higher value of λ_i, i.e. the one with lower expected cost. If the cost falls below a certain threshold ($\bar{c}$ in Equation (3.6)), the commuter takes that mode without observing the other one. If the cost of the first mode is higher than $\bar{c}$, the commuter observes the other mode. After observing both modes, the commuter takes the option with lowest cost.

When the commuter uses this strategy, his probability of choosing a multimodal trip is conditional on whether he chooses to observe multimodal options. He only does this if the observed cost of ride hailing is greater than the threshold for observing multimodal, $c_r \geq \bar{c}$. Therefore, the probability that the commuter will take a multimodal trip is

$$\begin{aligned} P(\text{multimodal}|\lambda_r, \lambda_m, s) =& P(c_m < c_r | c_r > \bar{c}) P(c_r > \bar{c}) \\ =& \left(\int_{\bar{c}}^{\infty} (1 - e^{-\lambda_m p}) \lambda_r e^{-\lambda_r p} dp\right) \left(\int_{\bar{c}}^{\infty} \lambda_r e^{-\lambda_r p} dp\right) \\ =& \frac{e^{-\bar{c}\lambda_r}\left(-e^{-\bar{c}(\lambda_r+\lambda_m)}\lambda_r + e^{-\bar{c}\lambda_r}(\lambda_r + \lambda_m)\right)}{(\lambda_r + \lambda_m)} \end{aligned} \tag{3.7}$$

Given this model, our data allows us to find relative values for λ_r, λ_m, and s. We observe the ratio of multimodal trips to pure ride-railing trips without a multimodal component, both before and after the introduction of an aggregator. After the aggregator is introduced, the ratio of trips will be purely determined by λ_r and λ_m according to Equation (3.4). The degree to which this differs from the pre-aggregator ratio of trips will be determined by the relative magnitude of the search cost, s.

[2]The Lambert W function with respect to x is the inverse of xe^x, so $W(xe^x) = x$ (see Corless et al., 1996).

3.2.1 Welfare

We calculate the welfare increase by finding the change in expected cost.

$$
\begin{aligned}
E(\text{cost w/o aggregator}) =& (s + E(c_r|c_r < \bar{c}))P(c_r < \bar{c}) \\
&+ P(c_r > \bar{c})[2s + E(c_m|c_m < \bar{c})P(c_m < \bar{c}) \qquad (3.8)\\
&+ E(min\{c_m, c_r\}|c_m > \bar{c}, c_r > \bar{c})P(c_m > \bar{c})] \\
=& s\left(\int_0^{\bar{c}} \lambda_r e^{-\lambda_r p} dp\right) + \int_0^{\bar{c}} p\lambda_r e^{-\lambda_r p} dp \\
&+ \left(\int_{\bar{c}}^{\infty} \lambda_r e^{-\lambda_r p} dp\right)\left[2s + \int_0^{\bar{c}} p\lambda_m e^{-\lambda_m p} dp\right] \qquad (3.9)\\
&+ \int_{\bar{c}}^{\infty} p(\lambda_r + \lambda_m) e^{-(\lambda_r+\lambda_m)p} dp \\
=& s(1 - e^{-\lambda_r \bar{c}}) + \frac{1 - \lambda_r \bar{c} e^{-\lambda_r \bar{c}} - e^{-\lambda_r \bar{c}}}{\lambda_r} \\
&+ e^{-\lambda_r \bar{c}}\left[2s + \frac{1 - \lambda_m \bar{c} e^{-\lambda_m \bar{c}} - e^{-\lambda_m \bar{c}}}{\lambda_m}\right] \qquad (3.10)\\
&+ \bar{c} e^{-(\lambda_r+\lambda_m)\bar{c}} + \frac{e^{-(\lambda_r+\lambda_m)\bar{c}}}{(\lambda_r + \lambda_m)}
\end{aligned}
$$

The expected cost with an aggregator is given in Equation (3.3). We revisit these equations in Section 3.6 to calibrate the model with data. Given the multimodal rate of travel with and without an aggregator, we can calculate relative values for λ_m, λ_r, and the relative change in expected cost when an aggregator is made available.

3.3 Data and Analysis

To understand how users changed their behaviour in response to the multimodal feature, we use a data set consisting of all ride-hailing trips booked through the aggregator app for the nine months before and the nine months after the treatment date, November 27th, 2018. There were 110,214 recorded trips in this time, 30,775 before the treatment and 79,439 after it. Uber accounts for 42% of all trips, Lyft accounts for 27%, and the other 31% of trips are spread out over various smaller services. In order to detect implicit multimodal behaviour, we use a data set of the locations of 891,325 transit stops from publicly available GTFS feeds.

We compare the start and end locations of all ride-hailing trips against transit stop locations using a k-nearest-neighbour search algorithm (Crookston and Finley, 2008).

On November 27th, 2018, a popular mapping app was updated to include a multimodal feature that allowed users to plan multimodal routes connecting ride hailing with rail. The feature also supported bus connections in nine cities with participating transit agencies. Trips are considered to connect with a transit stop if they start or end within one meter of a stop. Existing users are users who first installed the app prior to the introduction of the multimodal feature.

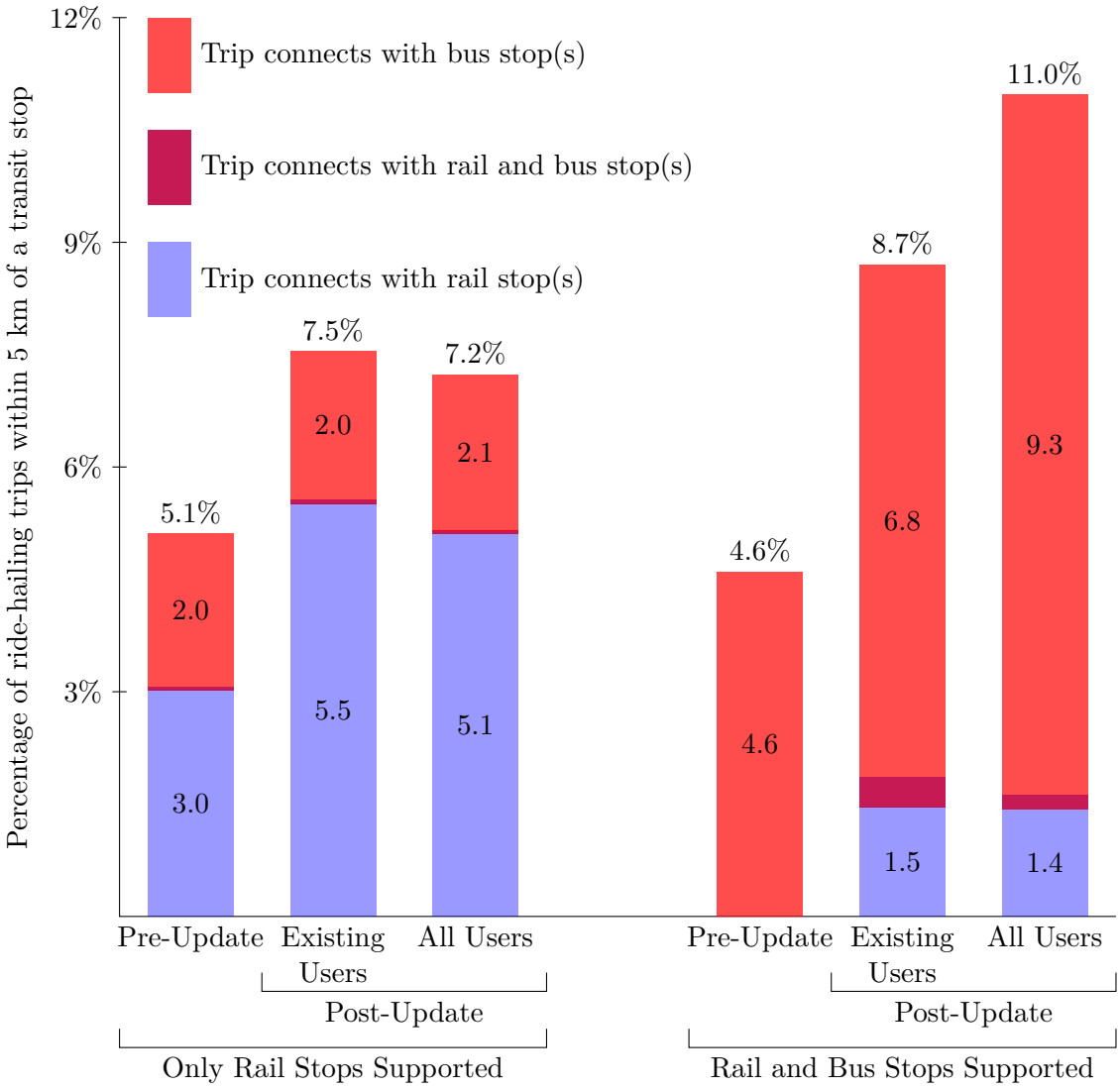


Figure 3.2 shows how the multimodal rate changed after the introduction of the multimodal trip planning service. The multimodal rate is defined as

$$\text{Multimodal rate} = \frac{\text{Rides starting or ending within 1 meter of a transit stop}}{\text{Rides starting or ending within 5 kilometers of a transit stop}}. \quad (3.11)$$

While the choices of 1 meter and 5 kilometers are arbitrary, the results are robust to changes in these specifications (see Appendix C.1).

After the introduction of the multimodal feature, the multimodal rate for rail stops increased from 3.0% to 5.1% ($p < 0.0001$, Fisher's exact test), as more users incorporated rail into their ride-hailing trips. Bus routes that were not included in the multimodal feature saw an insignificant change from 2.0% to 2.1% ($p = 0.5334$, Fisher's exact test).

In nine cities with participating transit agencies (Albany, Columbus, Dayton, Detroit, Kansas City, Las Vegas, Nashville, St. Louis, and St. Petersburg, FL) the app supported multimodal connections to bus routes. In these cities, bus routes saw their multimodal rate shoot up from 4.6% to 11.0% ($p < 0.0001$, Fisher's exact test). However, some of this increase can be attributed to a selection effect rather than a causal change on individual behaviour, as new users who may have been attracted to the app because of the multimodal feature tended to take more multimodal bus trips when possible. Among users who first downloaded the app prior to the introduction of the multimodal feature, the multimodal rate for supported bus routes had a smaller (but still significant, $p = 0.0020$, Fisher's exact test) increase from 4.6% to 8.7%.

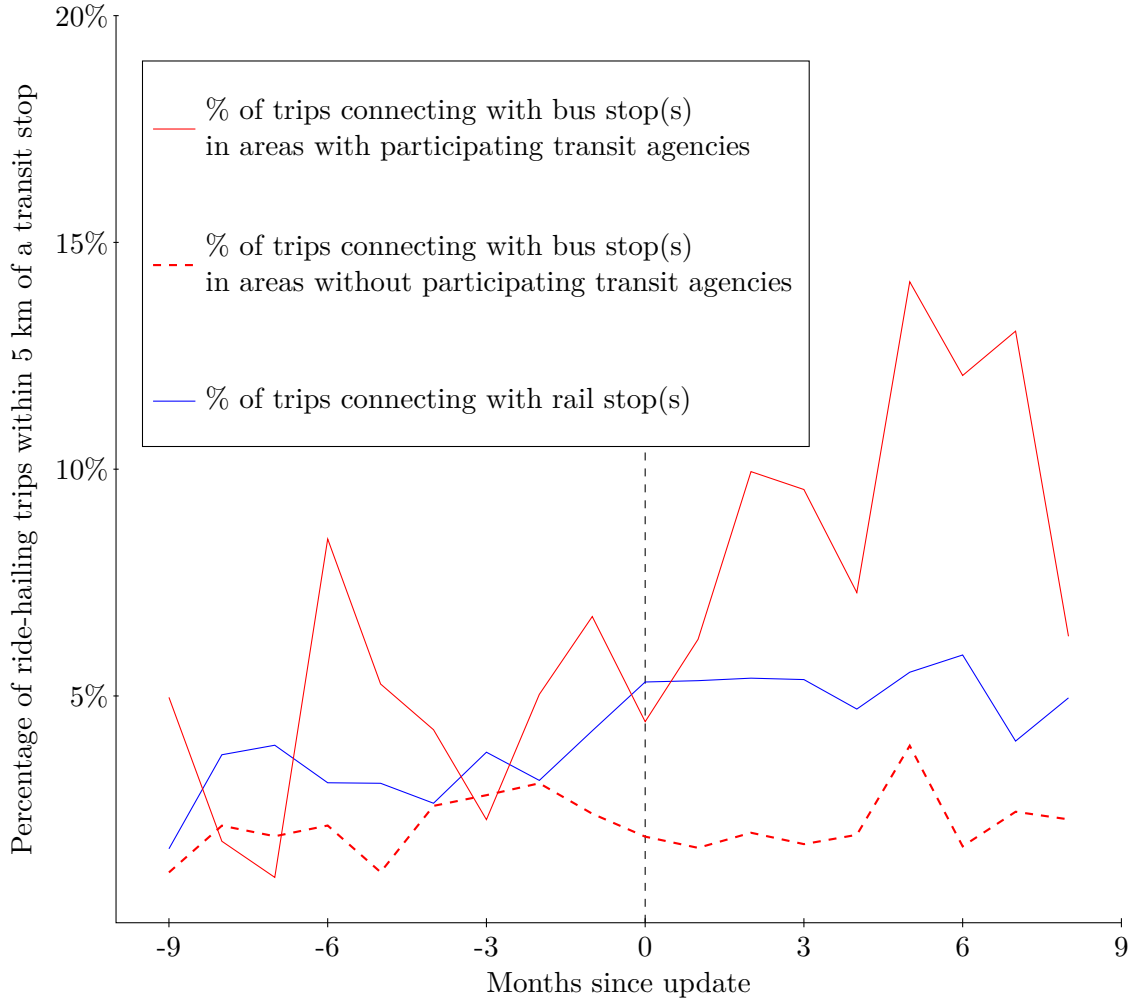
Figure 3.3 shows how these multimodal rates changed over time. These results are suggestive of a causal effect, with both rail and bus stops affected by the feature seeing a permanent increase in traffic after its introduction, and unaffected bus stops seeing no change. To verify this intuition, we develop a fixed-effects model that exploits the inclusion of bus trips in some areas but not others.

Transit agencies in nine cities in the United States collaborated with the app to allow users to plan multimodal trips connecting ride hailing to bus lines. Outside of the areas served by these nine agencies, users could only plan multimodal trips connecting to rail lines. We establish the causal impact of the multimodal bus feature by comparing the Combined Statistical Areas (CSAs) in the United States serviced by participating transit agencies to all other CSAs in the US with a sufficient volume of trips (dropping those with fewer than 50).

The participating transit agencies fall within nine CSAs: Albany-Schenectady in NY, Columbus-Marion-Zanesville in OH, Dayton-Springfield-Sidney in OH, Detroit-Warren-Ann Arbor in MI, Kansas City-Overland Park-Kansas City in MO and KS, Las Vegas-Henderson in NV and AZ, Nashville-Davidson-Murfreesboro in TN, St. Louis-St. Charles-Farmington in MO and IL, and Tampa-St. Petersburg-Clearwater in FL. We refer to these CSAs as treated areas.

After dropping all CSAs with fewer than 50 total trips and trips that cannot be identified with a CSA, we are left with a total of 77468 trips that served by bus (i.e., whose origin or destination is within 5km of a bus stop) in 44 combined statistical area. Among them, 7523 trips are in the 9 treated statistical areas: 1234 before the treatment and 6289 after it. There are 69945 trips are in the 35 non-treated statistical areas: 14986 before the treatment and 54959 after it.

This figure shows the monthly multimodal rates of ride hailing trips from February 27, 2018 to August 14, 2019. On November 27, 2018, an update allowed users to plan multimodal trips. They could connect with rail lines everywhere, and with bus lines in select cities with participating transit agencies. Trips are considered to connect with a transit stop if they start or end within one meter of a stop.



3.4 Empirical Models

We study the effect of the multimodal feature on multimodal trips using two distinct models: a fixed-effects model and an event-study model. We focus on the effect of multimodal support for bus travel on trips connecting with bus stops, as the partial rollout of this feature allows for clean causal identification.

3.4.1 Fixed-Effects Model

Our fixed-effects model exploits the difference in multimodal rates before and after the treatment day. Our specification is presented in Equation (3.12), where a refers to the CSA and t refers to the number of months since the treatment date.

$$Y_{ita} = \gamma_0 + \gamma_1 T_{ta} + area_a + month_t + \varepsilon_{ita} \tag{3.12}$$

The dependent variable Y_{ita} is a binary variable indicating whether trip i's origin or destination is within 1 meter of a bus stop in area a and t months since treatment. The variable of interests is the indicator variable T_{ta}. $T_{ta} = 1$ if the trip is within the treated areas and it happened after the introduction of multimodal planning services. Otherwise, $T_{ta} = 0$. $area_a$ and $month_t$ are the area and month fixed effects to control for the locations and seasonality.

Since the dependent variable is binary, we use both logit and probit models in our analysis. γ_1 measures the treatment effects.

3.4.2 Event Study Model

Our event-study model focuses on the aggregate weekly rate of multimodal trips. It is presented in Equation (3.13), where τ is the number of weeks since the treatment date and α is a binary variable denoting all trips in the CSAs supporting bus trips (when $\alpha = 1$) and in all other regions (when $\alpha = 0$).

$$y_{\tau\alpha} = \beta_0 + \beta_1 T_{\tau\alpha} + \beta_2 x_{\tau\alpha} + \beta_3 \tau + \varepsilon_{\tau\alpha} \tag{3.13}$$

The dependent variable, $y_{\tau\alpha}$, is the rate of trips in week τ and area α that start or end within 1 meter of a bus stop. $T_{\tau\alpha} = 1$ when $\tau >= 0$ and $\alpha = 1$. $x_{\tau\alpha}$ is the rate of trips ending within 1 meter of a rail stop. The model also includes a linear time trend.

3.5 Results

Our fixed-effects model results are presented in Table 3.1. The coefficient on T, our variable of interest, is consistently positive and significant at the 5% significance level across all specifications. In the logit model including month and area fixed effects, the result implies that after the introduction of multimodal trip planning, the log odds of the multimodal bus rate has increased by 0.721. So taking the initial multimodal rate of 4.6% in the treated areas from Figure 3.2, this implies that the multimodal feature caused an increase of 4.4%. Counterfactually, the model implies that rolling out the multimodal feature for bus routes in the non-treated areas would increase bus connections from 2.1% of ride-hailing trips to 4.2%.

Table 3.1: Effects of multimodal trip planning on bus connected ride-hailing trips

Dependent Variable: Whether a trip's origin or destination is within 1m of a bus stop						
	Logit Model			Probit Model		
	1	2	3	4	5	6
Multimodal trip planning	0.717 (0.231)*** [0.200]***	0.769 (0.226)*** [0.203]***	0.780 (0.270)*** [0.206]***	0.353 (0.101)*** [0.090]***	0.376 (0.099)*** [0.091]***	0.391 (0.121)*** [0.095]***
Month Fixed effects	N	Y	Y	N	Y	Y
Area Fixed effects	N	N	Y	N	N	Y

Notes: The dependent variable is an indicator whether variable showing whether a trip is multimodal, i.e., its origin or destination is within 1m of a bus stop. Standard errors clustered at the month level are in parentheses and Newey-West standard errors are in brackets. All the trips' origin or destination in the sample is within 5km of a bus stop. Areas with less than 50 trips are excluded.
* Statistical significance at the 0.1 level.
** Statistical significance at the 0.05 level.
*** Statistical significance at the 0.01 level.

When we limit the analysis to existing users only, the results are no longer significant when month and area fixed effects are included, as shown in Table 3.2. The effect size falls from 0.721 to 0.463 in the logit model, implying that part of the measured effect came from attracting new users who were more interested in multimodal travel than the existing ones.

Table 3.3 shows the results of the event study model from Equation (3.13). Consistent with the fixed-effects model, it shows the introduction of the multimodal feature causing a positive and significant increase in the rate of ride hailing trips connecting with bus stops, ranging from a 3.6 to a 6.3 percentage point increase across specifications.

3.6 Discussion

Mobility as a Service is a new industry. Its success depends on the cooperation of policy makers and incumbent mobility operators such as Uber and Lyft. This study has shown that, by providing multimodal trip planning services alongside the features of a conventional mapping app, a MaaS app can successfully encourage consumers to use combined routes involving ride hailing and public transit.

Only 3.1% of trips before the addition of the multimodal route planning feature connected to a rail stop. This increased to 5.6% among existing users after the addition of the feature (Figure 3.2). Referring back to our model from Section 3.2, we can estimate how high search costs must be relative to the cost distributions of ride-hailing and multimodal trips. Setting the parameter for ride-hailing $\lambda_r = 1$, a multimodal rate of 5.6% in the absence of search costs implies $\lambda_m = 0.059$ (Equation (3.4)). This in turn implies a threshold value of $\bar{c}(s, \lambda_m) = 0.498$ (Equation (3.7)) and a search cost of $s = 0.007$ (Equation (3.6)). This

Table 3.2: Effects of multimodal trip planning on existing users

Dependent Variable: Whether a trip's origin or destination is within 1m of a bus stop						
	Logit Model			Probit Model		
	1	2	3	4	5	6
Multimodal trip planning	0.600 (0.240)** [0.243]**	0.649 (0.228)*** [0.246]***	0.468 (0.298) [0.274]*	0.277 (0.108)** [0.112]**	0.295 (0.104)*** [0.113]***	0.248 (0.133)* [0.130]*
Month Fixed effects	N	Y	Y	N	Y	Y
Area Fixed effects	N	N	Y	N	N	Y

Notes: The dependent variable is an indicator whether variable showing whether a trip is multimodal, i.e., its origin or destination is within 1m of a bus stop. Standard errors clustered at the month level are in parentheses and Newey-West standard errors are in brackets. All the trips' origin or destination in the sample is within 5km of a bus stop. Areas with less than 50 trips are excluded. Only existing users' data is used. Existing users are users who first installed the app prior to the introduction of the multimodal feature.
* Statistical significance at the 0.1 level.
** Statistical significance at the 0.05 level.
*** Statistical significance at the 0.01 level.

Table 3.3: Event study model

	Dependent variable:			
	Bus rate			
	(1)	(2)	(3)	(4)
Rail rate		-0.408^{***}		-0.488^{***}
		(0.140)		(0.156)
Multimodal bus support	0.063^{***}	0.053^{***}	0.050^{***}	0.036^{***}
	(0.008)	(0.009)	(0.010)	(0.010)
τ	0.00002	0.0003	-0.00004	0.0003
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Constant	0.030^{***}	0.044^{***}	0.030^{***}	0.047^{***}
	(0.004)	(0.006)	(0.004)	(0.007)
Observations	154	154	154	154
Existing users only	N	N	Y	Y
Note:				*p<0.1; **p<0.05; ***p<0.01

relatively small search friction is sufficient to cause 2.5% of travellers in our sample to take the more costly mode of transportation. The aggregator reduces costs both by reducing search costs and by preventing people from taking the more costly mode of transportation. Given our estimated parameters, the expected cost of a trip, inclusive of search costs, is 0.957 without an aggregator (Equation (3.10)), and 0.952 with one.

To contextualize these numbers, setting $\lambda_r = 1$ means that every other number is proportional to the average cost of a ride-hailing trip (inclusive of non-monetary costs). So if this average ride-hailing cost were \$20, the estimated fall in users' travel costs would be \$0.11 per trip. We leave it up to the reader to judge whether this is a large or small effect for a software update. Perhaps more interesting than the cost reduction is the implication that small decreases in search frictions can produce significant changes in travel patterns.

By incorporating rail and bus into trips, multimodal travel has lower unpriced externalities (congestion, pollution) than ride hailing. There is a public interest in promoting public transit ridership as an alternative to car travel in order to reduce pollution and congestion. MaaS can be a tool to that end by helping consumers connect the advantages of public transit with the flexibility and convenience of ride hailing.

Bibliography

Aleksandr Alekseev, Gary Charness, and Uri Gneezy. Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, 134:48–59, 2017.

Evangelia Anagnostopoulou, Jasna Urbančič, Efthimios Bothos, Babis Magoutas, Luka Bradesko, Johann Schrammel, and Gregoris Mentzas. From mobility patterns to behavioural change: leveraging travel behaviour and personality profiles to nudge for sustainable transportation. *Journal of Intelligent Information Systems*, 54(1):157–178, 2020.

Alan T. Arnholt and Ben Evans. *BSDA: Basic Statistics and Data Analysis*, 2017. URL `https://CRAN.R-project.org/package=BSDA`. R package version 1.2.0.

David Austen-Smith and Timothy J. Feddersen. Deliberation, preference uncertainty, and voting rules. *American Political Science Review*, 100(2):209â217, 2006.

John Manuel Barrios, Yael V. Hochberg, and Hanyi Yi. The cost of convenience: Ridesharing and traffic fatalities. *Working Paper*, 2019.

Leonardo J. Basso and Hugo E. Silva. Efficiency and substitutability of transit subsidies and other urban transport policies. *American Economic Journal: Economic Policy*, 6(4): 1–33, 2014.

Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2019. URL `https://CRAN.R-project.org/package=Matrix`. R package version 1.2-17.

Roland Bénabou. The economics of motivated beliefs. *Revue d 'économie politique*, 125(5): 665–685, 2015.

Sourav Bhattacharya, John Duffy, and Sun-Tak Kim. Compulsory versus voluntary voting: An experimental study. *Games and Economic Behavior*, 84:111–131, 2014.

Maria Bigoni and Davide Dragone. Effective and efficient experimental instructions. *Economics Letters*, 117(2):460–463, 2012.

Laurent Bouton, Aniol Llorente-Saguer, and Frédéric Malherbe. Unanimous rules in the laboratory. *Games and Economic Behavior*, 102:179–198, 2017.

Noli Brazil and David Kirk. Uber and metropolitan traffic fatalities in the united states. *American Journal of Epidemiology*, 184:kww062, 07 2016. doi: 10.1093/aje/kww062.

Markus K. Brunnermeier and Jonathan A. Parker. Optimal expectations. *American Economic Review*, 95(4):1092–1118, 2005.

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

Alessandra Cassar and Dan Friedman. *Economics lab: an intensive course in experimental economics*. Routledge, 2004.

Gary Charness, Aldo Rustichini, and Jeroen Van de Ven. Self-confidence and strategic behavior. *Experimental Economics*, 21(1):72–98, 2018.

Daniel L. Chen, Martin Schonger, and Chris Wickens. oTreeâan open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88 – 97, 2016.

Roy Chen, Yan Chen, and Yohanes Riyanto. Public knowledge in coordination games: Learning from non-replication. 2018.

Regina R. Clewlow and Gouri S. Mishra. Disruptive transportation: The adoption, utilitzation and impacts of ride-hailing in the united states. *UC Davis ITS Research Report*, 2017.

Katherine Baldiga Coffman. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660, 2014.

Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. Using big data to estimate consumer surplus: The case of uber. Technical report, National Bureau of Economic Research, 2016.

Jean Converse and Stanley Presser. *Survey Questions: Handcrafting the Standadized Questionnaire.* Sage, 1986.

Robert M. Corless, Gaston H. Gonnet, David E.G. Hare, David J. Jeffrey, and Donald E. Knuth. On the lambertw function. *Advances in Computational mathematics*, 5(1): 329–359, 1996.

Peter J Coughlan. In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting. *American Political science review*, 94(2):375–393, 2000.

Nicholas Crookston and Andrew Finley. yaimpute: An r package for knn imputation. *Journal of Statistical Software, Articles*, 23(10):1–16, 2008. ISSN 1548-7660. doi: 10.18637/jss.v023.i10. URL `https://www.jstatsoft.org/v023/i10`.

Douglas D. Davis and Charles A. Holt. *Experimental Economics*. Princeton University Press, 1993.

Gregory D. Erhardt, Sneha Roy, Drew Cooper, Bhargava Sana, Mei Chen, and Joe Castiglione. Do transportation network companies decrease or increase congestion? *Science advances*, 5(5):eaau2670, 2019.

Timothy Feddersen and Wolfgang Pesendorfer. Voting behavior and information aggregation in elections with private information. *Econometrica*, 65(5):1029–1058, 1997.

Timothy Feddersen and Wolfgang Pesendorfer. Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political science review*, 92(1): 23–35, 1998.

Timothy J. Feddersen and Wolfgang Pesendorfer. The swing voter's curse. *The American economic review*, pages 408–424, 1996.

Sebastian Fehrler and Niall Hughes. How transparency kills information aggregation: theory and experiment. *American Economic Journal: Microeconomics*, 10(1):181–209, 2018.

Lawrence E. Fouraker and Sidney Siegel. *Bargaining Behavior*. McGraw-Hill New York, 1963.

Daniel Friedman and Shyam Sunder. *Experimental methods: A primer for economists*. Cambridge University Press, 1994.

Dino Gerardi and Leeat Yariv. Deliberative voting. *Journal of Economic Theory*, 134(1): 317 – 338, 2007. ISSN 0022-0531.

David Gill and Victoria Prowse. A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503, 2012.

Jacob K. Goeree and Leeat Yariv. An experimental study of collective deliberation. *Econometrica*, 79(3):893–921, 2011.

Warwick Goodall, Tiffany Dovey, Justine Bornstein, and Brett Bonthron. The rise of mobility as a service. *Deloitte Rev*, 20:112–129, 2017.

Serena Guarnaschelli, Richard D. McKelvey, and Thomas R. Palfrey. An experimental study of jury decision rules. *The American Political Science Review*, 94(2):407–423, 2000.

Jonathan D. Hall, Craig Palsson, and Joseph Price. Is uber a substitute or complement for public transit? *Journal of Urban Economics*, 2018.

Robert Cornelius Hampshire, Chris Simek, Tayo Fabusuyi, Xuan Di, and Xi Chen. Measuring the impact of an unanticipated suspension of ride-sourcing in austin, texas. *SSRN Electronic Journal*, pages 1–20, 2017.

Lionel Henry and Hadley Wickham. *purrr: Functional Programming Tools*, 2019. URL `https://CRAN.R-project.org/package=purrr`. R package version 0.3.2.

Ralph Hertwig and Andreas Ortmann. Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24(3): 383–403, 2001.

Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*, 2018. URL `https://CRAN.R-project.org/package=stargazer`. R package version 5.2.2.

Charles A Holt and Susan K Laury. Risk aversion and incentive effects. *The American Economic Review*, 92(5):1644–1655, 2002.

Charles A. Holt and Susan K. Laury. Risk aversion and incentive effects: New data without order effects. *The American Economic Review*, 95(3):902–904, 2005.

Jonathan Yinhao Huang, Farhan Majid, and Mark Daku. Estimating effects of uber ride-sharing service on road traffic-related deaths in south africa: a quasi-experimental study. *J Epidemiol Community Health*, 73(3):263–271, 2019.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010. URL `http://imai.princeton.edu/research/mediation.html`.

Peraphan Jittrapirom, Valeria Caiati, A-M Feneri, Shima Ebrahimigharehbaghi, María J. Alonso González, and Jishnu Narayan. Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. *Urban Planning*, 2:13–25, 2017.

Edi Karni. A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606, 2009.

Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, pages 1–16, 2018.

David S. Kirk, Nicolo Cavalli, and Noli Brazil. The implications of ridehailing for risky driving and road accident injuries and fatalities. *Social Science & Medicine*, 250:112793, 2020.

Mark T. Le Quement and Isabel Marcin. Communication and voting in heterogeneous committees: An experimental study. *Journal of Economic Behavior & Organization*, 174: 449–468, 2020.

Leib Litman, Jonathan Robinson, and Tzvi Abberbock. Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2):433–442, Apr 2017.

Jacopo Magnani and Ryan Oprea. Why do people violate no-trade theorems? a diagnostic test. Working paper, 2017.

Microsoft and Steve Weston. *foreach: Provides Foreach Looping Construct for R*, 2017. URL `https://CRAN.R-project.org/package=foreach`. R package version 1.4.4.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

Matthew Rabin. Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5):1281–1292, 2000.

Abhijit Ramalingam, Antonio Morales, and James Walker. Instruction length and content: Effects on punishment behaviour in public goods games. *Journal of Behavioral and Experimental Economics*, 73:66–73, 2018.

David Robinson and Alex Hayes. *broom: Convert Statistical Analysis Objects into Tidy Tibbles*, 2019. URL `https://CRAN.R-project.org/package=broom`. R package version 0.5.2.

Deepayan Sarkar. *lattice: Trellis Graphics for R*, 2018. URL `https://CRAN.R-project.org/package=lattice`. R package version 0.20-38.

Vernon L. Smith. Microeconomic systems as an experimental science. *The American Economic Review*, 72(5):923–955, 1982.

Erik Snowberg and Leeat Yariv. Testing the waters: Behavior across participant pools. Technical report, National Bureau of Economic Research, 2018.

Jana Sochor, Helena Strömberg, and I. C. MariAnne Karlsson. Implementing mobility as a service: Challenges in integrating user, commercial, and societal perspectives. *Transportation Research Record*, 2536(1):1–9, 2015.

Jana Sochor, Hans Arby, IC MariAnne Karlsson, and Steven Sarasini. A topological approach to mobility as a service: A proposed tool for understanding requirements and effects, and for aiding the integration of societal goals. *Research in Transportation Business & Management*, 27:3–14, 2018.

Zachary Spicer, Gabriel Eidelman, and Austin Zwick. Patterns of local policy disruption: Regulatory responses to uber in ten north american cities. *Review of Policy Research*, 36 (2):146–167, 2019.

Till Tantau. *The TikZ and PGF Packages*, December 2013. URL `http://sourceforge.net/projects/pgf/`.

Jean-Robert Tyran. Voting when money and morals conflict: an experimental test of expressive voting. *Journal of Public Economics*, 88(7-8):1645–1664, 2004.

Jacob W Ward, Jeremy J Michalek, Inês L Azevedo, Constantine Samaras, and Pedro Ferreira. Effects of on-demand ridesourcing on vehicle ownership, fuel consumption, vehicle miles traveled, and emissions per capita in us states. *Transportation Research Part C: Emerging Technologies*, 108:289–301, 2019.

Hadley Wickham, Jim Hester, and Romain Francois. *readr: Read Rectangular Text Data*, 2018. URL `https://CRAN.R-project.org/package=readr`. R package version 1.3.1.

Hadley Wickham, Romain FranÃ§ois, Lionel Henry, and Kirill MÃŒller. *dplyr: A Grammar of Data Manipulation*, 2019. URL `https://CRAN.R-project.org/package=dplyr`. R package version 0.8.1.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2019. URL `https://CRAN.R-project.org/package=knitr`. R package version 1.24.

Appendix A

Supplementary Appendix to "Three Angry Men: A Framed Jury Experiment"

A.1 Using Real Trials

The treatment involving real criminal trials presented a problem: With the other treatments, the true state of the world was randomly generated and known to the experimenter, but real trials are by their nature uncertain. In order to meet the requirements of the experiment, I needed a selection of criminal trials and evidence that fit three requirements:

1. There must be a reliable, independent measure of the factual innocence or guilt of the accused.
2. The trial evidence must be available in some form to show to experimental subjects.
3. The set of innocent and guilty cases should not be systematically different from each other in ways that affect beliefs independently of the guilt or innocence of the accused.

It is in the nature of criminal trials for the guilt of the accused to be in doubt, so the first requirement, finding an independent measure of factual guilt, was crucial. The National Registry of Exonerations provides such a measure. As the registry's website states, "The Registry provides detailed information about every known exoneration in the United States since 1989—cases in which a person was wrongly convicted of a crime and later cleared of all the charges based on new evidence of innocence."[1]

This registry is a strong indication of factual innocence. Since people convicted of crimes no longer benefit from the presumption of innocence, their exoneration requires strong

[1]Source: Our Mission. The National Registry of Exonerations. `https://www.law.umich.edu/special/exoneration/Pages/mission.aspx`. Retrieved on 2019-10-11.

affirmative evidence; the burden of proof in such cases is much higher than the reasonable doubt required for acquittal in a criminal case.

The second requirement—that the trial evidence is available to show to the subjects—was a limiting factor since the evidence in criminal trials is not generally made public. Furthermore, the evidence had to be summarized to be evaluated within the time frame of the experiment. However, summarizing adds the possibility of introducing bias if the person summarizing knows the outcome.

To get evidence summaries without introducing bias, I relied on reports from journalists present during criminal trials. I limited cases to homicides since these are much more likely to receive media attention than other crimes. For the cases drawn from the National Registry of Exonerations, I sought contemporary articles written during and immediately after the initial trials. I slightly edited articles that were published immediately after trials to obscure the verdict since knowledge of the verdict could bias subjects' perceptions of the evidence. Otherwise, these articles were presented to subjects without alteration.

Next, I gathered the cases that were not drawn from the registry. These were cases where the defendant was convicted of a crime and not exonerated; this was my independent metric for factual guilt. In keeping with the third requirement—that innocent and guilty cases should not be systematically different from each other—these cases were drawn in matched pairs from the other articles written by the same journalists or news organizations that reported on the cases from the exoneration registry. For each case drawn from the registry and summarized by a journalist, the experiment included one other case summarized by the same journalist or news organization around the same time.

This controls for several possible sources of bias in the selection of cases. Since the same journalists or news organizations summarized the innocent and guilty cases, the process of summarizing the evidence affected both sets of cases independently. The use of contemporary reports makes the experiment effectively double-blind: Neither the journalists nor the experimental subjects knew who would eventually be exonerated. Since these journalists and news organizations are locally based, the matched pairs were drawn from the same legal jurisdictions, controlling for possible jurisdictional differences.

Finally, I restricted the cases based on their details. Experimental subjects had to assign subjective probabilities to the factual guilt or innocence of the accused. Therefore, there must be a clear factual question of guilt. Some court cases involve debates about legal culpability rather than debates about factual guilt, and these were excluded.

Name Changes The difficulty of doing online research and using real cases is that subjects may simply Google the names of people involved in cases. For this reason, I substituted their real names for fake ones. In order to avoid changing the implied ethnicities of people involved in cases, which may impact subjects' perceptions, I searched each name on `behindthename.com`, a website that provides the etymological origins of many names, and then I found a substitute name with a similar origin. A name like Garcia, with a Portuguese origin, was replaced by another Portuguese name like Cruz. A Welsh name like Maddox was substituted for another Welsh name such as Sayer. This process was somewhat subjective, but I expect it to have very little impact on the outcomes of the study.

A.2 Experimental Instructions

Consent for Participation in Experimental Research

Purpose: The purpose of this study, entitled "Deliberation," is to test different models of decision-making.

Description: This Human Intelligence Task (HIT) asks you to make a series of choices among alternatives that involve monetary prizes. The HIT should take approximately 20-30 minutes to complete. Your answers will be used in an academic study on decision-making.

Researchers:

Garrett M. Petersen (Principal Investigator)

PhD Candidate

Department of Economics

Simon Fraser University

Burnaby BC, Canada

[redacted]@sfu.ca

David Freeman

Assistant Professor

Department of Economics

Simon Fraser University

[redacted]@sfu.ca

Confidentiality: Your decisions will be kept strictly confidential. We have made every effort to guarantee your privacy and anonymity. Following the completion of the experiment, data will be stored on the SFU vault cloud storage service, and as such will be protected under the BC Freedom of Information and Protection of Privacy Act. You will never be identified by name or any other identifying feature with relation to this study. This experiment was created with oTree and is hosted on the Heroku cloud platform. Any data hosted within the United States may be subject to information requests from government agencies under the U.S. Patriot Act.

Compensation: You will be paid $2.00 USD for completing this HIT and an additional bonus payment between $0.10 and $1.50 may be made, depending on an element of chance and on your choices.

Potential Risks: This study will involve text-only interaction between you and other participants. We cannot guarantee the behaviour of your fellow participants, and as such this experiment carries the same risk of abusive behaviour as all anonymous online interactions.

Contact for Information about the Study: If you have any questions or desire additional information with respect to this study, you may contact Garrett M. Petersen at Tel: [redacted], or email: [redacted]@sfu.ca.

If you have any concerns about your rights as a research participant and/or your experiences while participating in this study, you may contact Dr. Jeffrey Toward, Director, SFU Research Ethics [redacted]@sfu.ca or [redacted].

Your participation in this study is entirely voluntary and you may refuse to participate or withdraw from the study at any time. You can print this form of the consent form and maintain it for your own records. NOTE: Please do not take this HIT if you are not willing to commit 30 minutes of your full concentration to the HIT. The data we collect is being used for scientific research. We greatly appreciate your full attention and careful consideration of each question.

This HIT (or any version of it) can only be taken once by each worker. If you complete this HIT more than once, you will only be paid for the first time.

To consent to participate in the study, enter your Worker ID in the field below, then click 'Next.'

[Text input box]

(Note: Your worker ID must match the one you used to sign up for this HIT.)

[Next button]

Instructions (1/4)

In this experiment, you will be assigned to a group of 3 to make a joint decision.

The experiment has a number of different scenarios, so we will give a general overview before your group is assigned and the scenario is determined.

Each scenario has two possible states of the world, e.g. "Red" and "Blue." One of these states will be randomly selected *with equal probability* prior to the start of the experiment. In the first stage of the experiment, you and your group members will each be shown evidence related to the true state.

After observing your evidence, you will be asked to make two choices: First, you will get to cast a vote for what you think the true state is. Second, you will be asked to state your belief as a percentage (0-100%) in one of the two states. Neither your votes nor your beliefs will be shown to other players at any time.

Click "Next" to proceed to the next page of the instructions.

[Next button]

Instructions (2/4)

Once you have cast your initial vote and stated your initial belief, you will proceed to the deliberation stage. This stage will feature a chat box where you may discuss anything you like with your fellow group members. The chat stage will last a fixed amount of time, depending on the scenario.

After you have completed the deliberation stage, your entire group will get the chance to vote again, and to report your updated beliefs, just like in the first stage. These are the final decisions in the experiment. After everyone has made their decisions, your payments will be determined.

The next page of the instructions will tell you how payments are determined. Click "Next" to proceed.

[Next button]

Payment (3/4)

To recap, by the end of the experiment you will have made four decisions:

1. An initial vote
2. An initial belief
3. A final vote
4. A final belief

One of these four will be randomly selected, with equal probability, to determine your payment.

If a vote is selected, then what matters is what the majority of group members voted for. If the majority voted correctly (i.e. for the true state determined at the start), then everyone will get a high payoff. (Payoffs vary slightly in different scenarios, so we will tell you exactly what they are once your group has been assigned and the scenario has been selected.) If the majority voted incorrectly, then everyone in the group will get a low payoff of $0.10.

The next page describes how payments are determined if a belief is selected.

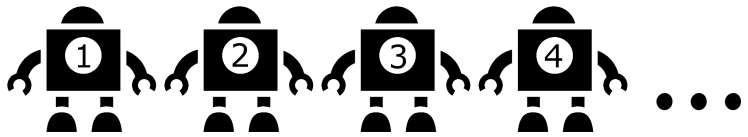
[Next button]

Payment (4/4)

If a belief is selected for payment, each group member's payoff will be separately determined by their own belief.

The mechanism for determining payoffs from a belief are as follows:

Imagine 100 robots numbered from 1 to 100, all lined up in order.



Each of these robots is capable of predicting the true state of the world with a probability equal to their number. For instance, Robot 14 is only capable of predicting correctly 14 times out of every hundred, whereas the much cleverer Robot 89 is capable of predicting correctly 89 times out of every hundred. Robot 50 is no better or worse at predicting than a coin toss.

Let's say you are trying to guess whether the true state of the world is "Heads" or "Tails." You don't have any evidence one way or the other, so when asked to state your belief that the state is "Heads," you say your belief is 50%.

If this belief is selected for payment, a random robot will be selected from the line. If the robot has a higher number than your belief (68 for instance), then the robot will guess for you, and you will earn $1.00 if he is correct and $0.10 otherwise. If the robot has a lower or equal number than your belief (41 for instance) then the robot won't matter and you will get $1.00 if the true state is "Heads" and $0.10 otherwise.

You maximize your chance of getting a high payoff by truthfully reporting what you believe. If your true belief is 50% and you say it's 55%, then you will miss out on the chance to have robots 51 to 54 guess for you, and they all have better odds than you do. If you say your belief is 45%, then you might have robots 46 to 49 guess for you, and they all have worse odds than you do.

[Next button]

Summary and Quiz

The following points summarize the instructions:

- You will be sorted into a group of 3.
- A true state of the world will be selected at random by the computer.
- You will be shown evidence pointing towards what that true state is.
- You and your group members will vote on the true state.

- You will report your belief, as a percentage chance, in one of the possible true states.
- Your group will deliberate for five minutes.
- You and your group members will vote and state your beliefs a second time.
- One vote or belief will be selected for payment. Each of the four choices (first vote, first belief, second vote, second belief) have an equal chance of being selected.
- If a vote is selected, the whole group gets a high payoff if a majority voted correctly.
- If a belief is selected, all players are paid individually according to the robot mechanism:
  - A random robot from 1 to 100 is selected.
  - If the robot's number is less than or equal to your belief, you get a high payoff so long as the outcome being predicted is true.
  - If the robot's number is greater than your belief, the robot guesses for you.
  - Robot number X guesses correctly X percent of the time.

Quiz

Complete the following quiz to proceed to the experiment.

1. In the deliberation phase, why should you care what other players think?

   ○ Because their beliefs matter for your payoff.
   ◉ Because their votes matter for your payoff.
   ○ You shouldn't care.
   ○ Because truth is a value onto itself.

2. You are asked for your belief that the sun will rise tomorrow. Assuming you are certain that it will, what belief will maximize your odds of a high payoff under the robot mechanism?

   [Slider from 0 to 100]

3. You state a belief of 25% that two coin flips will both come up heads. Robot 37 is selected. What happens?

   ○ The robot is ignored. You get a high payoff if at least one coin is not heads.
   ○ The robot is ignored. You get a high payoff if both coins come up heads.
   ○ The robot guesses for you. It has a 25% chance of guessing correctly.
   ◉ The robot guesses for you. It has a 37% chance of guessing correctly.

4. You state a belief of 25% that two coin flips will both come up heads. Robot 15 is selected. What happens?

○ The robot is ignored. You get a high payoff if at least one coin is not heads.

◉ The robot is ignored. You get a high payoff if both coins come up heads.

○ The robot guesses for you. It has a 25% chance of guessing correctly.

○ The robot guesses for you. It has a 15% chance of guessing correctly.

5. What are the odds that the first vote will be selected for payment?

○ 50%

○ 100%

◉ 25%

○ 33%

6. Do you have any pressing appointments in the next half hour or so?

◉ No, I don't

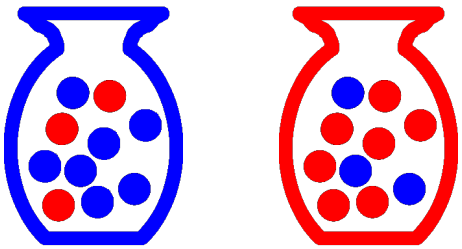
○ Actually...

Warning: If you get more than 20 wrong quiz answers through successive tries on the quiz, you will be prevented from proceeding, and your HIT will be rejected.

[Next button]

Scenario [Objective Nonpartisan Treatment]

There are two jars: one red and one blue. The red jar contains 7 red balls and 3 blue balls, while the blue jar contains 3 red balls and 7 blue balls.



The computer has randomly selected one of these jars, each having an equal chance of being the one selected.

On the next screen, you will be shown one ball that has been randomly selected from the true jar. Your group members will each be shown a different draw (with replacement) from the jar. You will be casting votes and selecting beliefs on the colour of the true jar.

If one of the votes is selected for payment, then your group will earn $1.00 each if at least two members voted for the true jar colour, and $0.10 otherwise.

If one of the beliefs is selected for payment, you will be individually paid according to the robot mechanism described previously. Remember, you maximize your chances of getting a high payoff by always truthfully reporting your beliefs.

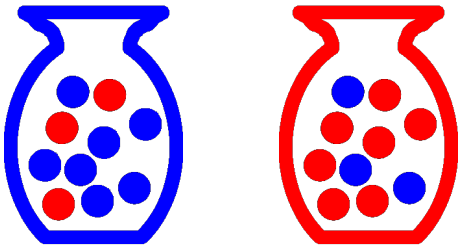
[Next button]

The following table shows the bonus payments in all possible cases. It will be displayed below later pages in case you need to review it.

Vote Selected for Payment		
	Colour is Red	Colour is Blue
Majority Votes Red	\$1.00	\$0.10
Majority Votes Blue	\$0.10	\$1.00
Belief Selected for Payment		
	Colour is Red	Colour is Blue
Belief > Robot Number	\$1.00	\$0.10
Belief ≤ Robot Number	\$1.00 if robot guesses right, \$0.10 if robot guesses wrong	

Scenario [Objective Partisan Treatment]

There are two jars: one red and one blue. The red jar contains 7 red balls and 3 blue balls, while the blue jar contains 3 red balls and 7 blue balls.



The computer has randomly selected one of these jars, each having an equal chance of being the one selected.

On the next screen, you will be shown one ball that has been randomly selected from the true jar. Your group members will each be shown a different draw (with replacement) from the jar. You will be casting votes and selecting beliefs on the colour of the true jar.

If one of the votes is selected for payment, then your group will earn only \$0.10 each if a majority vote for the wrong jar colour. If a majority votes correctly, then things are a little more complicated. Each member of your group has an equal chance of "leaning" towards one colour or the other. If the group votes for the answer you lean towards, and that answer is correct, you get \$1.50. If the group votes for the answer you don't lean towards, and that answer is correct, you get only \$0.50. It is still always better to choose the right colour, but if the right colour is the colour you lean towards, that's even better for you.

You lean towards RED.

If one of the beliefs is selected for payment, you will be individually paid according to the robot mechanism described previously. Remember, you maximize your chances of getting a high payoff by always truthfully reporting your beliefs.

[Next button]

The following table shows the bonus payments in all possible cases. It will be displayed below later pages in case you need to review it.

Vote Selected for Payment		
	Colour is Red	Colour is Blue
Majority Votes Red	\$1.50	\$0.10
Majority Votes Blue	\$0.10	\$0.50
Belief Selected for Payment		
	Colour is Red	Colour is Blue
Belief $>$ Robot Number	\$1.00	\$0.10
Belief $\leq$ Robot Number	\$1.00 if robot guesses right, \$0.10 if robot guesses wrong	

Scenario [Subjective Nonpartisan Treatment]

The computer has randomly picked a colour: red or blue. Both colours had an equal chance of being selected.

On the next screen, you and the other members of your group will all be given exactly 10 seconds to look at the same 20 by 20 grid of red and blue balls. There will be exactly ten more balls of the colour the computer picked.

You will be casting votes and selecting beliefs on which colour has more balls in the grid.

If one of the votes is selected for payment, then your group will earn \$1.00 each if at least two members voted for the correct colour, and \$0.10 otherwise.

If one of the beliefs is selected for payment, you will be individually paid according to the robot mechanism described previously. Remember, you maximize your chances of getting a high payoff by always truthfully reporting your beliefs.

[Next button]

The following table shows the bonus payments in all possible cases. It will be displayed below later pages in case you need to review it.

Vote Selected for Payment		
	Colour is Red	Colour is Blue
Majority Votes Red	\$1.00	\$0.10
Majority Votes Blue	\$0.10	\$1.00
Belief Selected for Payment		
	Colour is Red	Colour is Blue
Belief > Robot Number	\$1.00	\$0.10
Belief ≤ Robot Number	\$1.00 if robot guesses right, \$0.10 if robot guesses wrong	

Scenario [Subjective Partisan Treatment]

The computer has randomly picked a colour: red or blue. Both colours had an equal chance of being selected.

On the next screen, you and the other members of your group will all be given exactly 10 seconds to look at the same 20 by 20 grid of red and blue balls. There will be exactly ten more balls of the colour the computer picked.

You will be casting votes and selecting beliefs on which colour has more balls in the grid.

If one of the votes is selected for payment, then your group will earn only \$0.10 each if a majority vote for the wrong colour. If a majority votes correctly, then things are a little more complicated. Each member of your group "leans" towards one colour. If the group votes for the answer you lean towards, and that answer is correct, you get \$1.50. If the group votes for the answer you don't lean towards, and that answer is correct, you get only \$0.50. It is still always better to choose the right colour, but if the right colour is the colour you lean towards, that's even better for you.

You lean towards BLUE.

If one of the beliefs is selected for payment, you will be individually paid according to the robot mechanism described previously. Remember, you maximize your chances of getting a high payoff by always truthfully reporting your beliefs.

[Next button]

The following table shows the bonus payments in all possible cases. It will be displayed below later pages in case you need to review it.

Vote Selected for Payment		
	Colour is Red	Colour is Blue
Majority Votes Red	\$0.50	\$0.10
Majority Votes Blue	\$0.10	\$1.50
Belief Selected for Payment		
	Colour is Red	Colour is Blue
Belief > Robot Number	\$1.00	\$0.10
Belief ≤ Robot Number	\$1.00 if robot guesses right, \$0.10 if robot guesses wrong	

Scenario [Framed Treatment]

On the next page, you will be shown a real news article about a jury trial. The article has been lightly edited, but no details of the evidence have been changed. It was written at the time of the trial by a journalist who was there in the court room. You and your group will be deciding whether the defendant is GUILTY or INNOCENT.

We ask that you **not** do any independent research. Please limit your investigation to the details of the article presented on the following page.

This case was randomly selected from a large number of cases. Importantly, in half of the cases in our set, later evidence emerged proving that the defendant could not have committed the crime, and they were exonerated of all legal consequences. These defendants are considered to be innocent for the purpose of this study. The other half of cases feature defendants who were convicted and whose convictions were never overturned. These defendants are considered to be guilty for the purpose of this study.

If one of the votes is selected for payment, then your group will earn $1.00 each if at least two members voted correctly, and $0.10 otherwise.

If one of the beliefs is selected for payment, you will be individually paid according to the robot mechanism described previously. Remember, you maximize your chances of getting a high payoff by always truthfully reporting your beliefs.

Since this version of the experiment takes longer than average, you will be given an additional bonus of $0.50 in addition to your other earnings throughout the experiment.

WARNING: The story on the following page may contain descriptions of crime, violence, and death. If you are easily disturbed, it is not recommended that you continue. Remember that you can exit the experiment at any point without penalty.

[Next button]

The following table shows the bonus payments in all possible cases. It will be displayed below later pages in case you need to review it.

Vote Selected for Payment		
	Colour is Red	Colour is Blue
Majority Votes Red	$1.50	$0.60
Majority Votes Blue	$0.60	$1.50
Belief Selected for Payment		
	Colour is Red	Colour is Blue
Belief > Robot Number	$1.50	$0.60
Belief ≤ Robot Number	$1.50 if robot guesses right, $0.60 if robot guesses wrong	

Appendix B

Supplementary Appendix to "Instructions" (Freeman, Kimbrough, Petersen, and Tong 2018)

B.1 Review of current practice

Inclusion/Exclusion criteria

We included experimental papers published between January 2011 and December 2016 in six journals: the American Economic Review, Econometrica, the Quarterly Journal of Economics, the Journal of Political Economy, the Review of Economic Studies, and Experimental Economics. Articles from the AER: Papers and Proceedings were excluded. In order to be included, a paper had to include at least one lab experiment. We excluded field experiments and online experiments that were not conducted in a controlled environment, but we include "lab-in-the-field" experiments that were conducted in a controlled environment.

To classify each included experiment, we reviewed both the text of each paper and supplementary materials available online through the journal's website, with the exception of uncompiled code (e.g. z-Tree code).

Coding Criteria: Delivery

Delivery methods could include paper instructions or computer instructions. Values in the supplementary table are 1 for yes, 0 for no, 0.5 for uncertain. In some cases, an alternative delivery method was used; for example, Etang *et al.* (2011) studied subjects in rural Cameroon and used purely verbal instructions because many subjects were illiterate.

We code the study as having paper instructions if it is directly stated or clearly implied that a set of paper instructions were used. Some papers were explicit about their use of printed instructions, while others required us to infer the existence of paper instructions from other details. For instance, Mittone and Ploner (2011, p. 207) write that "after the choices are collected, instructions for the beliefs elicitation phase are distributed." Distribution implies a written set of instructions, though this is not explicitly stated. Sometimes we inferred the form of instructions from the instructions themselves, for instance in Altmann *et al.* (2014), the instructions included screenshots, from which we inferred that they must have been printed on paper.

We code the study as having computer instructions if it is directly stated or clearly implied that computerized instructions were used. Sometimes this was explicit, while other times it had to be inferred. For instance, in papers that included copies of their instructions online, some instructions told participants to click on something to proceed to the next screen. This implies that the instructions are computerized, even if it is not explicitly stated in the text of that paper. Cox and James (2012, Supplement p. 2) end their instructions by telling their subjects, "When you have finished reading and have asked any questions you might have, please click Done."

Many papers are unclear on whether the instructions are given on paper or on computers. If there was no explicit statement of the form of instructions in the paper itself, and no clear indication from the instructions where these were available online, the paper was coded as uncertain.

Coding Criteria: Reinforcement

We coded four different forms of reinforcement.

1. Read aloud. We code an experiment as having read aloud its instructions if it is stated or clearly implied that the instructions were presented orally. Most often this meant that the experimenter read the instructions for the participants to hear. Some studies, such as Aycinena *et al.* (2014, p. 110), included voice recordings of the instructions, which we coded as read aloud as indicated by the following quote "They were provided with instructions and were also shown a video which read these instructions aloud."

2. Demonstration or guided practice. We code a paper as including demonstration or guided practice if we can infer that it used walk-throughs of the experimental interface, examples, or demonstrations of aspects of the experiment during the instructions phase. Walk-throughs involve actively-guided practice by the subject. Examples include hypothetical descriptions of potential actions and consequent outcomes. For instance, Brookins and Ryvkin (2014) give subjects an example of the likelihood of success, conditional on the group members' investment. Demonstrations actively highlight one or more aspects of the experiment, for example, throwing a die to show subjects how uncertainty will be resolved as in Ericson and Fuster (2011). The mere use of graphical or tabular methods to communicate information, or providing screenshots in paper instructions, was considered neither demonstration nor guided practice.

3. Unguided practice. If the experiment included one or more unpaid practice rounds without guidance, we coded this as unguided practice. Sometimes this was explicit in the body of the paper, while other times it was only indicated in the instructions themselves.

4. Quiz. Quizzes or questionnaires were only included if they occurred after the instructions and before the experiment. Many experiments include questionnaires to check participants' understanding ex post, but these are not counted as they do not reinforce participants' understanding of the instructions before the experiment.

When a quiz was given, we checked whether feedback was given after the quiz and before the experiment. If it was clearly stated that subjects were given the correct answers to the quiz, "Feedback" was coded as a 1. If subjects must get 100% to proceed with the experiment, we infer that feedback was given. Many papers give quizzes to "ensure comprehension of instructions" but do not explicitly indicate whether answers were given. For example Cabrera *et al.* (2013, p. 432) indicate that "subjects completed a quiz to make sure they had fully understood the logic of the game." It is ambiguous whether this implies that feedback was given to promote subject understanding ex-ante or instead quiz performance was used by the experimenters to assess subject comprehension ex-post. Such papers are coded as uncertain with respect to quiz feedback. We also separately code whether subjects were paid for correct quiz answers (Incentivized) and whether participants were required to get all questions correct before continuing (Require 100%).

Coding Criteria: Some main task(s) is (are) one shot

We classified the main task or tasks for each experiment. If at least one of the main tasks is one shot (that is, subject can be viewed as making a single decision) in one or more of the treatments, we coded that paper as having a one shot main task under this column. When researchers use a choice list or the strategy method – where multiple similar decisions are made almost simultaneously, and could in-principle be viewed as one decision – we view this task as a one-shot task. In contrast, when decisions are made in a sequence, even without feedback, we would not consider those to constitute a one-shot task. Anderson *et al.*'s (2011) study provides an edge case. In their experiment, each subject plays six public goods games with different parameter values, but all six choices are presented at the same time. Since all choices are instances of the same basic task and are presented at once, we coded their experiment as one shot. If these tasks had been presented sequentially on separate screens, we would not have coded this as one shot. An interesting boundary case is a dynamic game with an evolving state variable (e.g. the money supply variable in Petersen and Winn (2014)); subjects in such games make repeated decisions in the same task, but with different incentives depending on the state. We have coded these as repeated (i.e. not one shot) because there is typically feedback between decisions and the state dependence is usually not so severe that subsequent decisions differ fundamentally from those made in initial round. The opportunities for learning from repetition thus usually dominate (though not necessarily always), and we note that we did not explicitly account for this in our coding.

Coding Criteria: Some main task(s) has (have) feedback between decisions

If at least one of the main tasks was repeated with feedback between rounds in one or more of the treatments, we coded that paper as having a repeated main task with feedback under this column (e.g. a repeated public goods game in which subjects learned their payoff after each round (e.g. Bayer *et al.* (2013)). We considered it sufficient for a subsequent round to involve choices in the same basic task as the preceding one for which feedback was given. For example, in Noussair and Stoop (2015), subjects in one treatment completed two dictator games in a row, with different reward media (money and time) with feedback between them – we viewed these as repetitions of the same task with feedback.

Coding Criteria: More than one task

We coded whether an experiment has more than one incentivized task. In some cases, an experiment required subjects to input multiple separate decisions associated with the same broader task – in these instances, we coded this a single task (as discussed above). Sometimes a single task has multiple decisions (e.g. a centipede game as in Cox and James (2012) or a public goods game with punishment as in Harris *et al.* (2015)). Similarly, in an experiment that required subjects to vote on a sanctioning scheme that would then be implemented in a public goods game (Kamei *et al.*, 2015), we viewed the vote and the subsequent game as one task. Many experiments coded as having more than one task would follow up a main task with a secondary preference elicitation.

Cross-Check

Each paper was independently coded by two coders, who read each of the 260 papers in the review along with any instructions available in their online supplementary materials. For each of the 11 categories coded, both coders marked them as true (=1), false (=0), or uncertain (=0.5). Both coders agreed most of the time, only disagreeing (including cases where one coder was uncertain) in 363 out of 11×260 judgments, and only disagreeing fundamentally (i.e. one coder marking a "0" and the other a "1" on a given paper-category judgment) in 200 such judgments. The area with the most disagreement was the presence of demonstration, examples, or guided practice. These are particularly difficult to identify, as they are often buried in lengthy instructions and the difference between explanation and demonstration is somewhat subjective. We note that false negatives are more likely than false positives – it is easy to miss an example or demonstration in instructions but hard to see one where it doesn't actually exist. After each person coded independently, both coders reconciled disagreements to put together the data for Table 2.1. Typically, when only one coder was uncertain, disagreement was resolved in favour of the certain coder. In the case of genuine disagreement coders discussed and settled on the most likely classification.

Correlations amongst practices

One-shot experiments account for about one third of the experiments using computerized instructions (31%) or paper instructions (35%). 57% of experiments that use neither paper

Table B.1: Correlation between experiment type and delivery and reinforcement

	One-shot	p-value	Feedback between decisions	p-value
Paper only	.048	.437	.008	.899
Computer only	-.011	.863	-.082	.189
Both	.018	.770	.022	.722
Neither	.157	.011	-.180	.004
Read aloud	.112	.072	-.092	.141
Practice/Demonstration	-.191	.002	.190	.002
Quiz	-.146	.019	.159	.010
Table reports pairwise correlations between delivery/reinforcement category (rows) and experiment type (columns) and their p-values.				

Table B.2: Instruction practices by feedback

	One-shot	Feedback between decisions
Total	84	152
Read aloud	52	76
Practice/Demonstration	36	98
Quiz	24	69

nor on-screen instructions are one-shot games; most of these studies are field experiments in which experimenters read instructions aloud or go through the instruction one-on-one with subjects.

We also find that one-shot experiments tend to be less likely to use each of the reinforcement methods (except for reading aloud) – even though such experiments give no feedback, making each subject's initial understanding of the instructions crucial. We suspect that this is because one-shot experiments tend to be simpler and therefore easier to explain. Instructions are read aloud more often in one-shot game experiments (62%) than in experiments with feedback between decisions (50%). Other reinforcement methods are used less often in one-shot experiments than in experiments with feedback between decisions (respectively, 43% versus 65% use some form of practice or demonstration, while 29% versus 45% use a quiz). These differences result in a significant negative association between one-shot experiments and use of practice/demonstration ($\rho = -.191$, $p = .002$) and quizzes ($\rho = -.146$, $p = .019$) in the instructions.

B.2 Experimental Instructions

The experimental sessions all followed the script in Figure B.1.

Figure B.1: Experimenter's script for running a session

How to Run a Session

1. Log in to computer 24 with your SFU email
2. Log in to students' computers using username "econ subject" and password "economics" (computers 11 and 12 sometimes freeze!)
3. Open ESILauncher on computer 24
4. Highlight the machine numbers students are using
5. Check the Auto Connect box
6. Select the file "C:\Experiments\PoodleJump\Client\Client.exe"
    a. Replace leading dots with "C:\Experiments"
7. Open "C:\Experiments\PoodleJump\Server\Server.exe" on computer 24
8. Hit "Load Settings" button and select "C:\Experiments\PoodleJump\Server\ExperimentSettings\Low.txt"
9. As participants arrive, mark them as "participated" on http://experiments.econ.sfu.ca/
10. Set the number of participants in both ESI and Server
11. Give consent forms and receipts and instruct participants to fill out everything except the payment amount
12. Take in consent forms
13. Give the pre-experiment speech
    a. Eyes on own screen
    b. Don't communicate with other participants
    c. Raise hand to ask question
    d. No food
    e. Keep drinks in closed containers
    f. Cell phones away
    g. If doing paid quiz, explain about the paid quiz
14. Click the big green check mark in ESI to launch the program
15. Instruct subjects to click "Run"
16. Tell participants to sit quietly once they have finished instructions
17. (if doing quiz) Tell them about quiz (and incentives if quiz is incentivized)
18. Click "Begin Instructions"
19. Allow them to go through the instructions
20. (if doing quiz) Hand out quiz
21. (if doing quiz) Take in quiz
22. (if doing quiz + answers) Read quiz answers
23. Click start button
24. (if doing quiz) Grade quiz during the experiment
25. Mark experiment as "Finished" on http://experiments.econ.sfu.ca/
26. When experiment is complete, ask students to wait at their computers and have their receipts ready
27. Call students by computer number and pay them $7+their experiment payoff, filling out dollar amounts in each receipt
28. Move data files from "..\PoodleJump\Server\Server_Data\" into "Dropbox\PoodleJump\data\[appropriate folder]\"

We include copies of all instructions pages as seen by each subject in all treatments. First, we show the screenshots that apply for all except for the ENHANCED treatment. Note that the printed instructions for the paper treatment did not include the screenshots shown in

Figure B.5.4 and Figure B.7.6, since they completed practice periods for Tasks 1 and 2 as part of the on-screen instructions, like all other subjects.

Figure B.2: Instructions page 1: introduction to the experiment

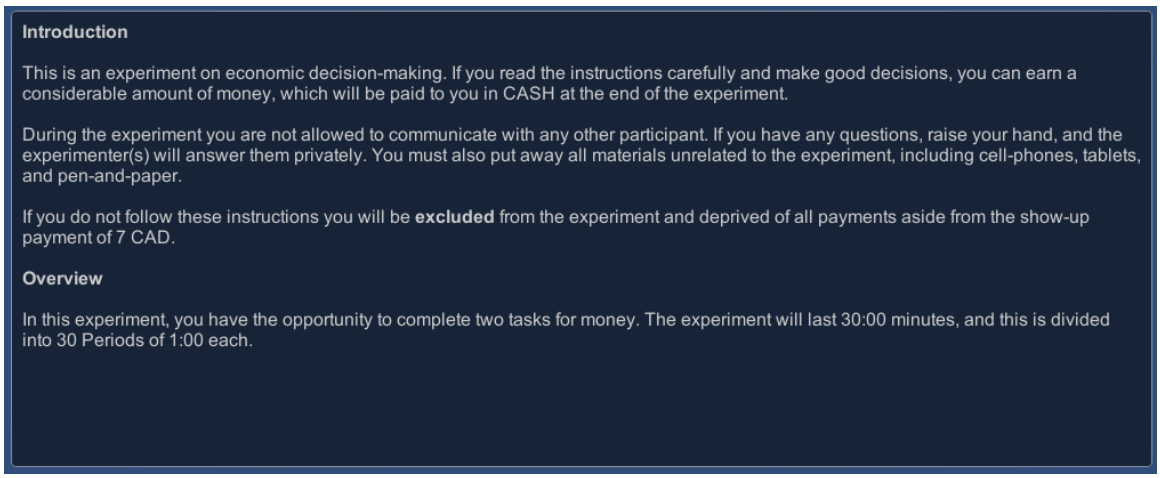


Figure B.3: Instructions page 2: description of Task 1

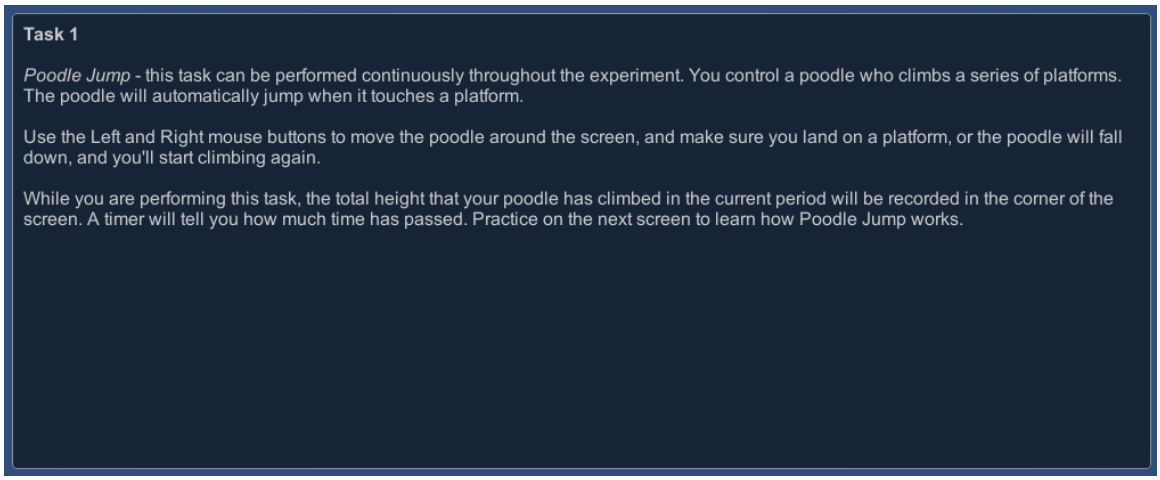


Figure B.4: Instructions page 3: Task 1 practice

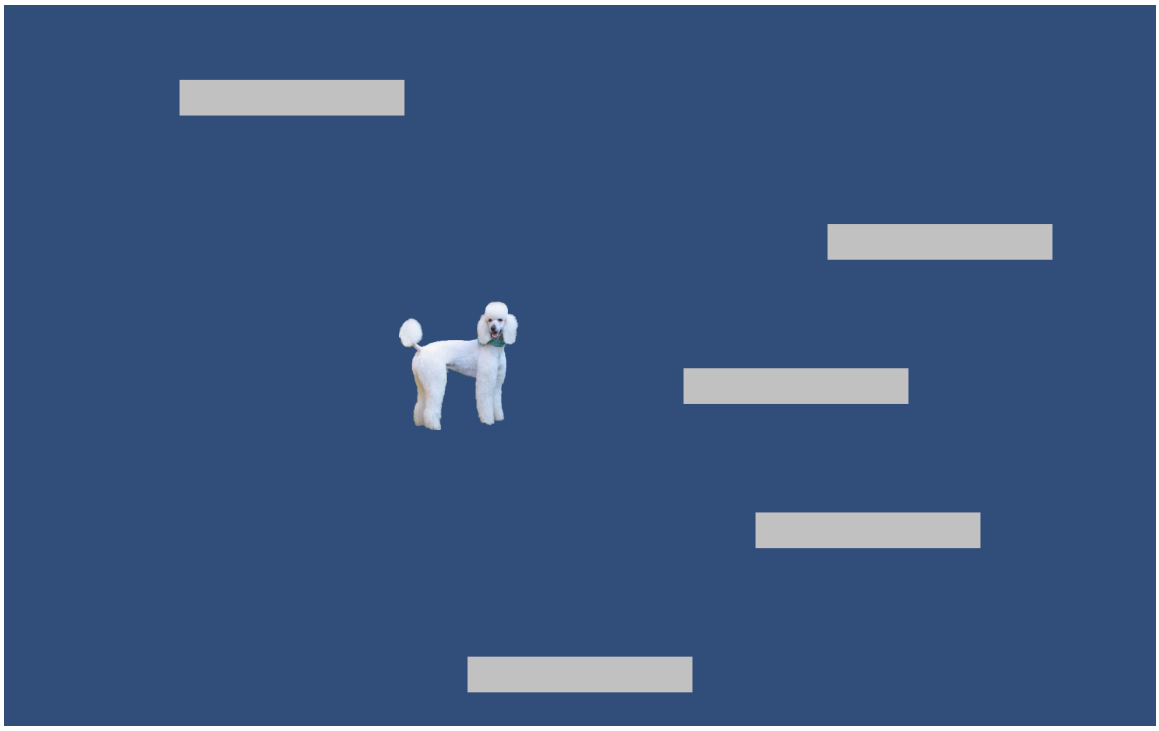


Figure B.5: Instructions page 4: description of Task 2

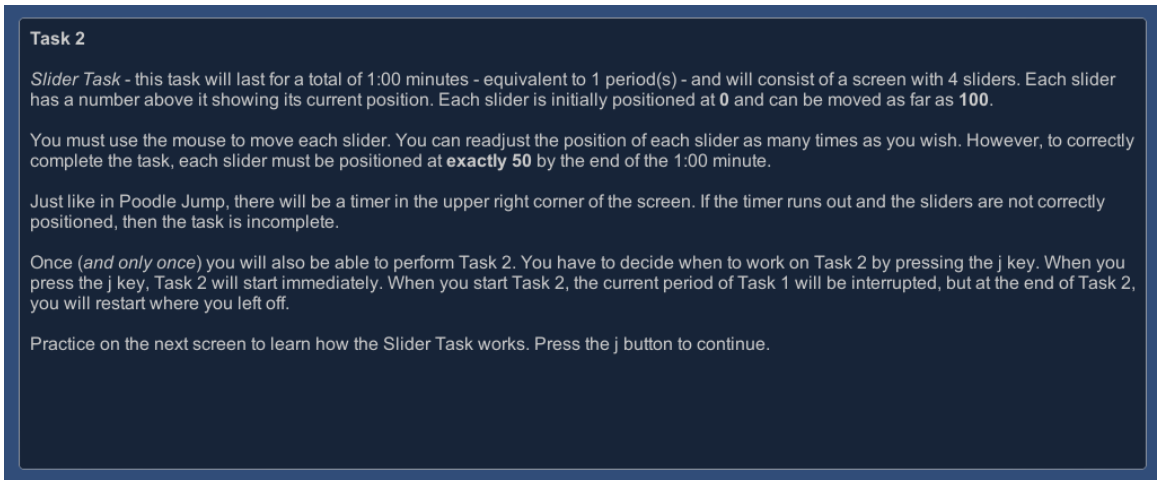


Figure B.6: Instructions page 5: Task 2 practice

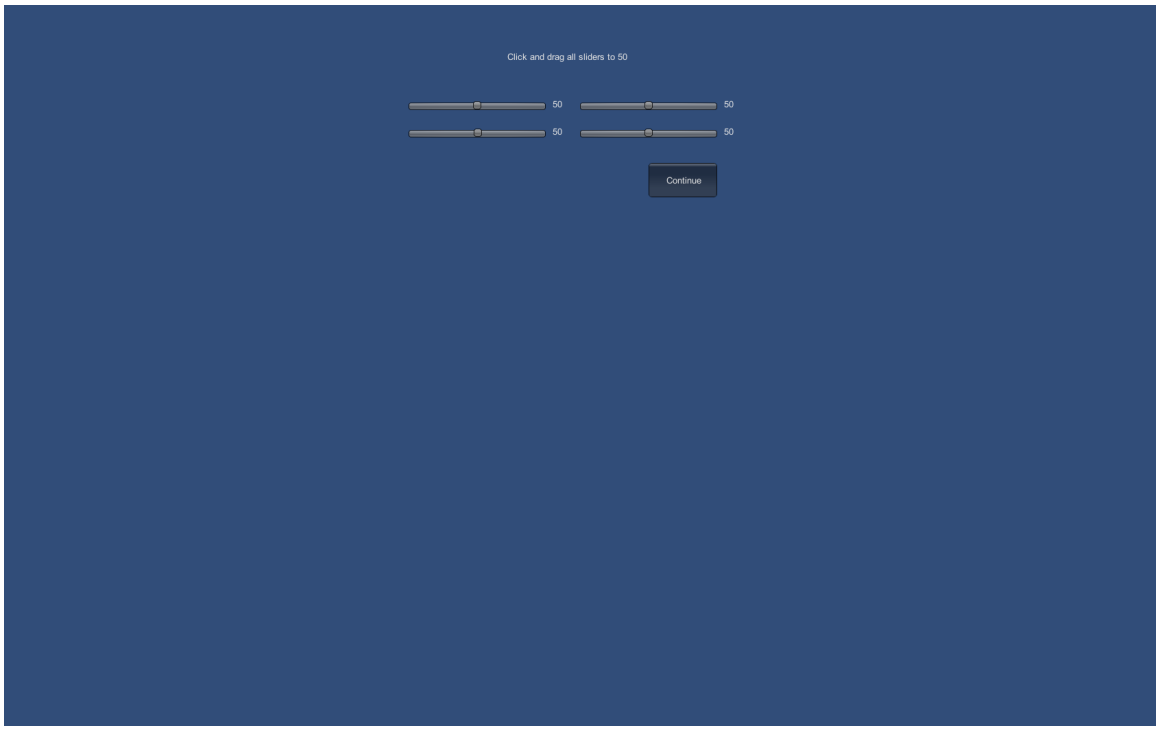


Figure B.7: Instructions page 6: payment schedule description

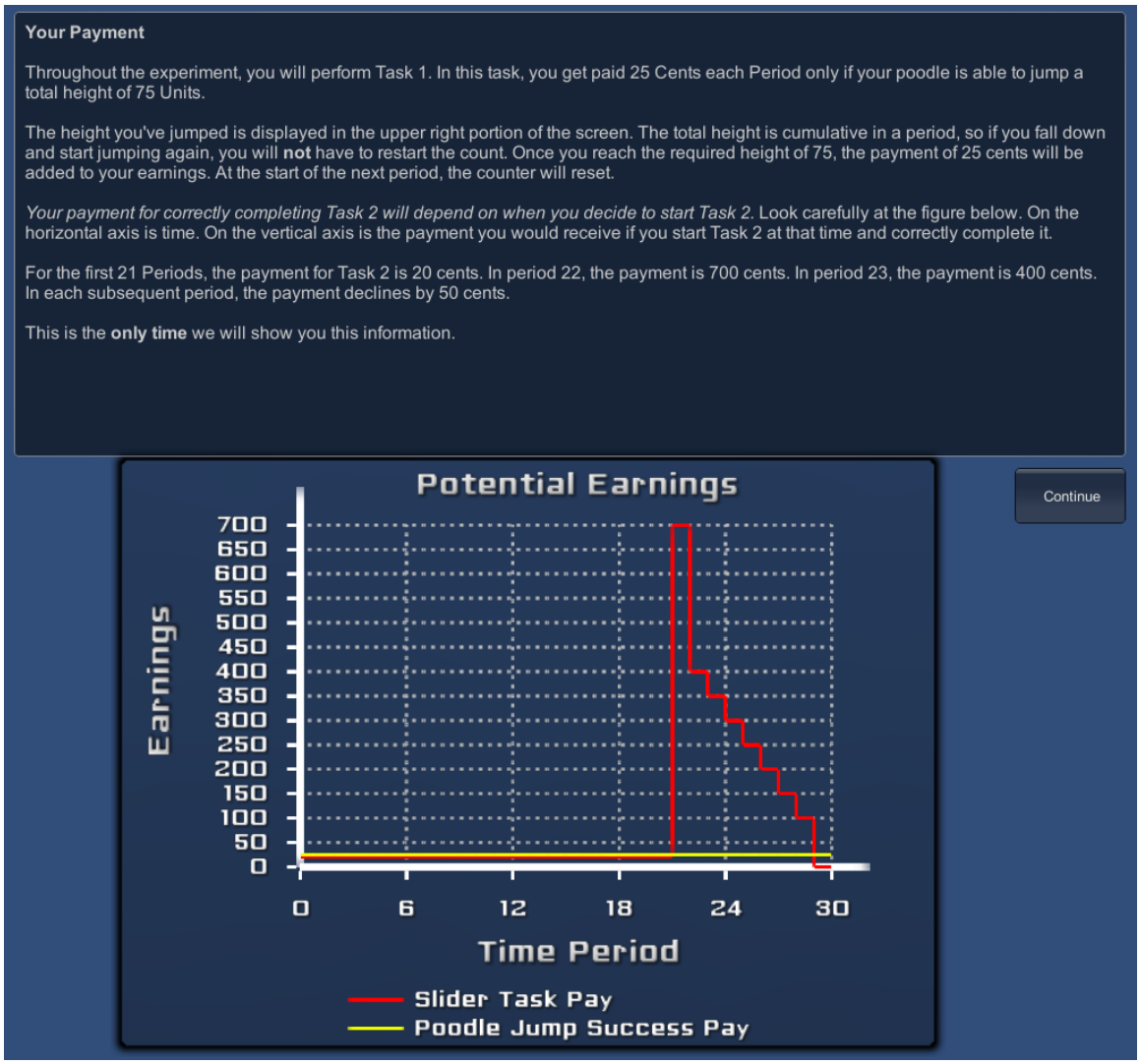
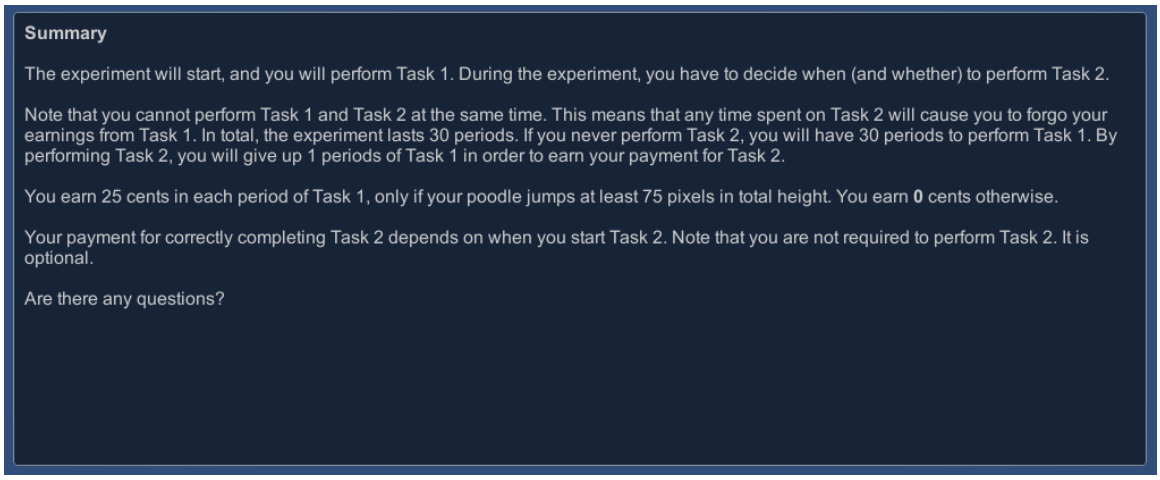


Figure B.8: Instructions page 7: summary



Next, we include screenshots from the instructions from the ENHANCED treatment. Note that, unlike in the other treatments, the final summary screen remained displayed in the ENHANCED while subjects wrote the quiz.

Figure B.9: ENHANCED Instructions page 1: introduction to the experiment

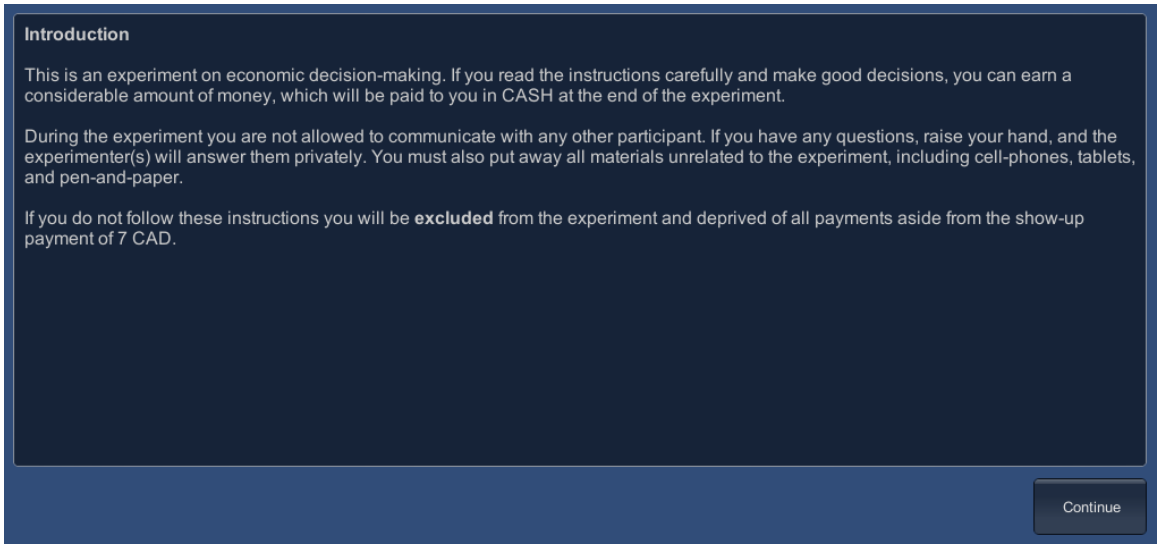


Figure B.10: ENHANCED Instructions page 2: overview and payment

Figure B.11: ENHANCED Instructions page 3: payment examples

Figure B.12: ENHANCED Instructions page 4: description of Task 1

Figure B.13: ENHANCED Instructions page 5: Task 1 practice

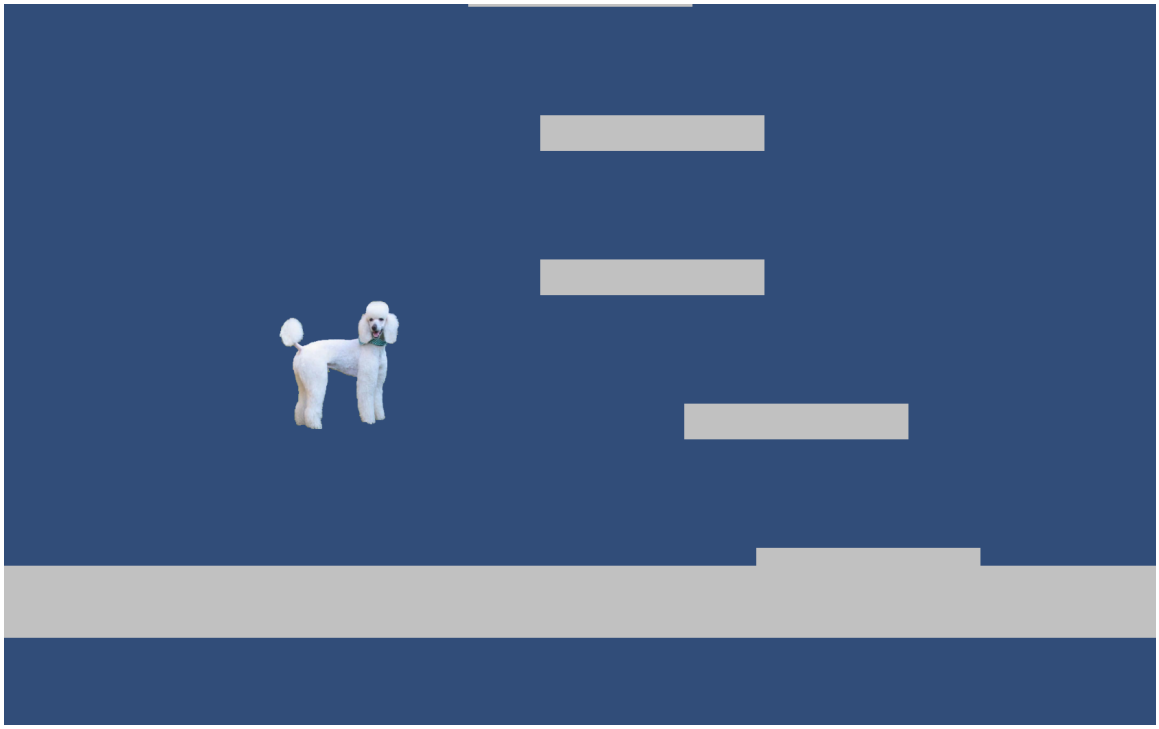


Figure B.14: ENHANCED Instructions page 6: description of Task 2

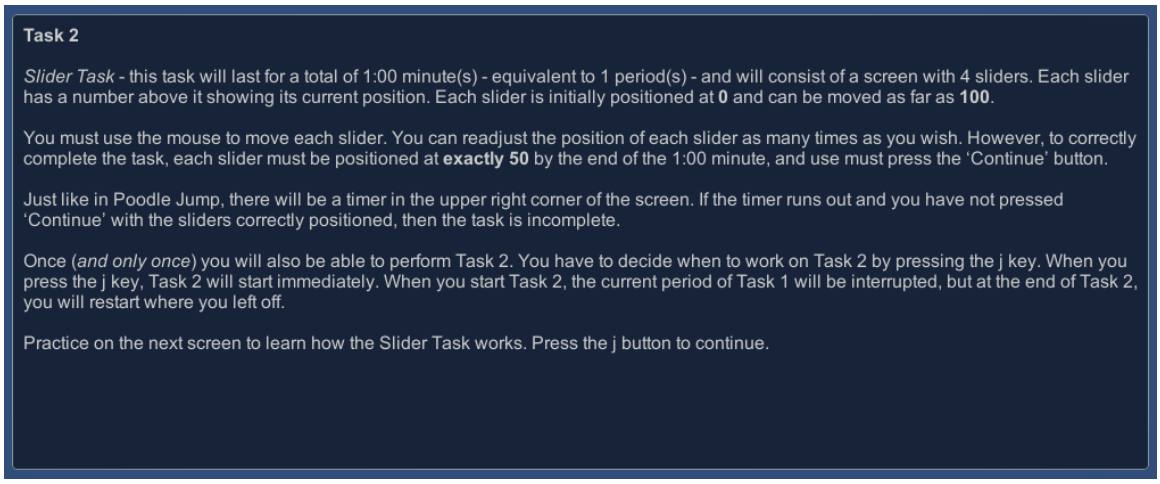


Figure B.15: ENHANCED Instructions page 7: Task 2 practice

Figure B.16: ENHANCED Instructions page 8: payment recap

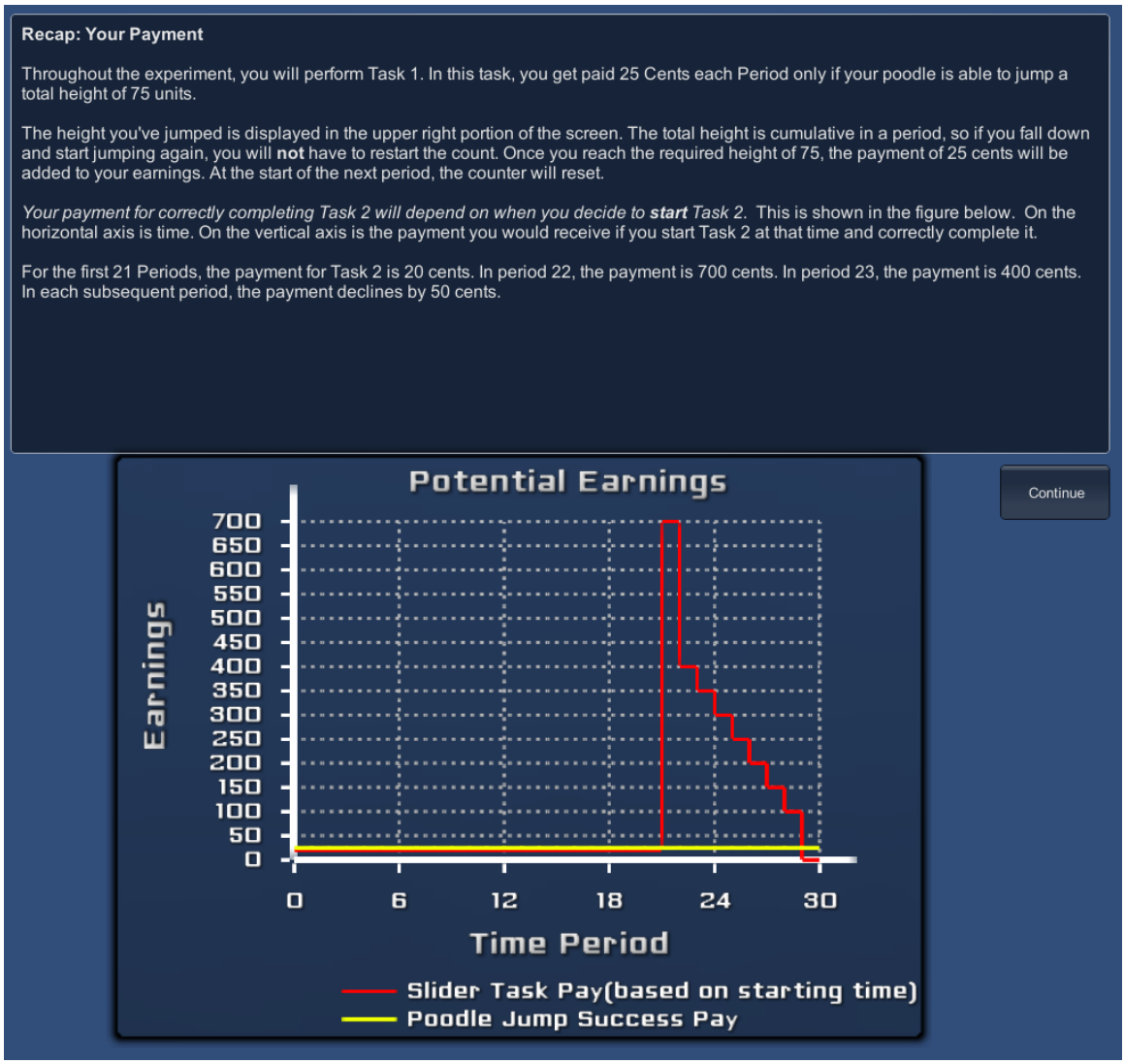
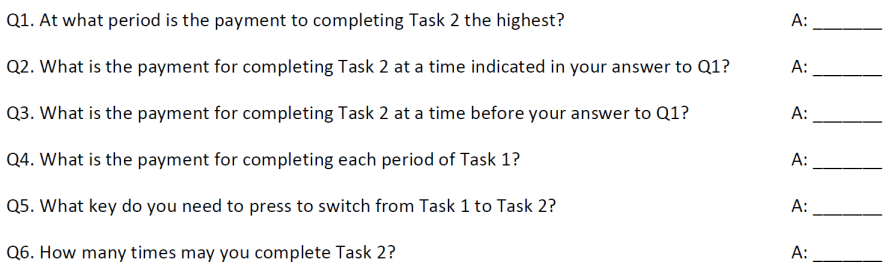


Figure B.17: ENHANCED Instructions page 7: summary

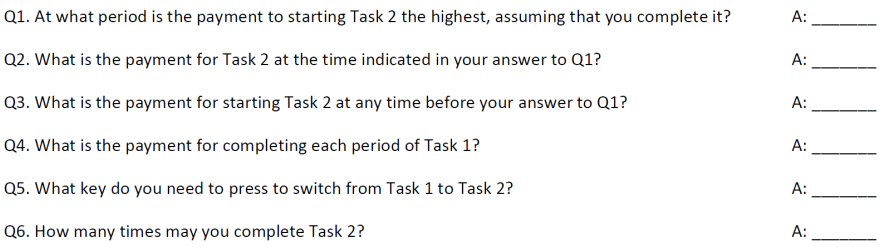
Our quiz, which was included after the instructions and before the main experiment in all treatments except for NO QUIZ, featured the following six questions:

Figure B.18: Post-instructions quiz



In our follow-up experimental sessions, we slightly re-worded some of the quiz questions to make them more clear. This new quiz was administered to all subjects in the ENHANCED treatment and some of the subjects in the QUIZ treatment.

Figure B.19: Revised post-instructions quiz

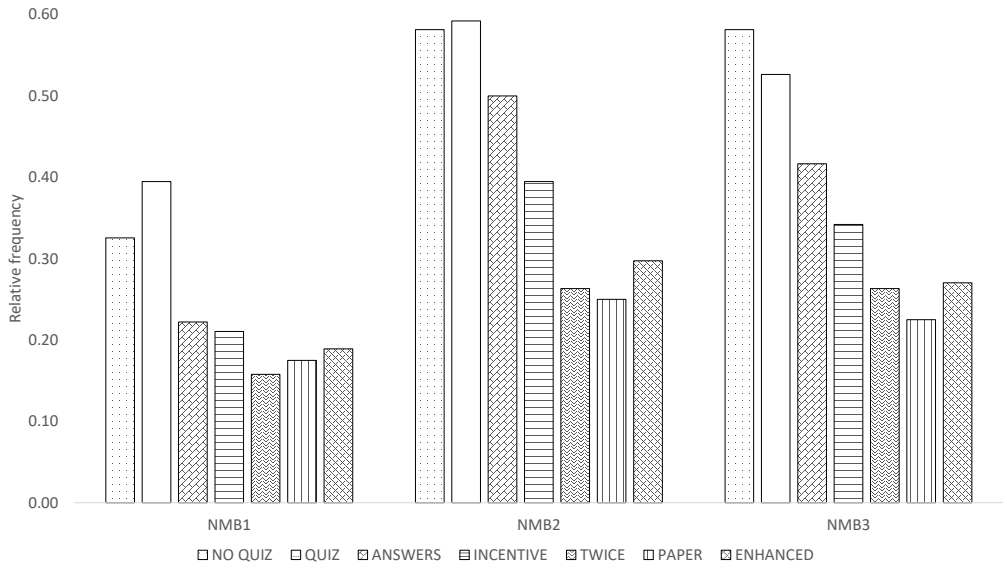


While scores in the QUIZ treatment did increase slightly under the new quiz, from an average of 3.9 to 4.4, this difference is not statistically significant ($p = .11$, rank-sum test), and thus we pool data from all QUIZ sessions. We also did not observe any significant differences in NMB ($p = .50$, Fisher's exact test).

B.3 Robustness checks

We redo our analysis with three alternative measures of NMB to check the robustness our results. The specifications reported in Table B.3.1-3 are all analogous to the specifications in Table 2.4, but with alternative definitions of NMB. The dependent variable "NMB1" is equal to one if the subject did Task 2 before period 21 and equal to zero otherwise; this measure of NMB allows for trembles. The "NMB2" variable defines any behavioural deviation from optimality as NMB. That is, it classifies a subject as exhibiting NMB unless they did Task 2 exactly in period 22. Finally, the "NMB3" variable classifies those who did Task 2 before period 22 or never at all as NMB. The results of these alternative specifications are broadly consistent with those reported in Table 4. Figure B.20.1 plots the share of subjects with NMB in each treatment, by each of these alternative measures. To check the robustness of our logit regressions, Table B.20.4 reports estimated linear probability models with (OLS analogues to columns 1 and 2 of Table 2.4); for comparison purposes note that we do not report marginal effects in Table 2.4 since the mediation analysis in column 4 provides the economically meaningful estimates of interest.

Figure B.20: Percentage of subjects revealing NMB, under three alternative definitions of NMB, by treatment.



We note that our statistical tests find significant differences between our main QUIZ treatment and each of our INCENTIVE, TWICE, PAPER, and ENHANCED treatments, but do not detect significant differences among the latter four treatments, and also detects no significant difference between the ANSWERS treatment and other treatments (see Table 2.2 in the main text). This raises the question of statistical power. We note that the comparisons between the QUIZ treatment and each of the INCENTIVE, TWICE, PAPER, and ENHANCED treatments appear to be appropriately powered. Across the latter four treatments, 21.6% of subject misunderstand (a fraction which ranges between 18.4-23.7%

Table B.3: Treatment effects on NMB1 and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB1		Quiz Score	NMB1	
	(1)	(2)	(3)	(4)	
NO QUIZ	-0.301				n
	(-1.096, 0.495)				
ANSWERS	-0.825*	0.207	-0.050	-0.169*	
	(-1.746, 0.096)	(-2.648, 3.061)	(-0.586, 0.487)	(-0.329, 0.008)	112
ANSWERS × Quiz Score		-0.324		0.005	
		(-1.062, 0.413)		(-0.056, 0.070)	
INCENTIVE	-0.894*	-1.380	0.211	-0.164*	
	(-1.810, 0.022)	(-4.202, 1.422)	(-0.354, 0.775)	(-0.331, 0.021)	114
INCENTIVE × Quiz Score		0.127		-0.022	
		(-0.508, 0.762)		(-0.091, 0.039)	
TWICE	-1.247**	-0.677	0.421	-0.199**	
	(-2.244, -0.249)	(-3.940, 2.586)	(-0.156, 0.998)	(-0.367, -0.010)	114
TWICE × Quiz Score		-0.135		-0.044	
		(-0.847, 0.578)		(-0.119, 0.016)	
PAPER	-1.123**	7.787**	1.320***	0.163	
	(-2.070, -0.176)	(1.053, 14.521)	(0.922, 1.718)	(-0.118, 0.375)	116
PAPER × Quiz Score		-1.632**		-0.133***	
		(-2.901, -0.363)		(-0.223, -0.046)	
ENHANCED	-1.028**	0.249	0.489*	-0.139*	
	(-1.981, -0.074)	(-3.675, 4.174)	(-0.030, 1.008)	(-0.325, 0.060)	113
ENHANCED × Quiz Score		-0.273		-0.051*	
		(-1.144, 0.598)		(-0.123, 0.003)	
Quiz Score		-0.519***			
		(-0.875, -0.164)			
Intercept	-0.427*	1.662**	4.105***		
	(-0.893, 0.038)	(0.121, 3.202)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

Table B.4: Treatment effects on NMB2 and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB2		Quiz Score	NMB2	
	(1)	(2)	(3)	(4)	
NO QUIZ	-0.044				n
	(-0.812, 0.724)				
ANSWERS	-0.373	-1.477	-0.050	-0.103	
	(-1.179, 0.434)	(-5.163, 2.209)	(-0.586, 0.487)	(-0.268, 0.070)	112
ANSWERS × Quiz Score		0.192		0.009	
		(-0.624, 1.009)		(-0.098, 0.116)	
INCENTIVE	-0.800*	-2.840	0.211	-0.164*	
	(-1.605, 0.004)	(-6.591, 0.911)	(-0.354, 0.775)	(-0.344, 0.013)	114
INCENTIVE × Quiz Score		0.443		-0.042	
		(-0.353, 1.239)		(-0.157, 0.069)	
TWICE	-1.402***	-2.201	0.421	-0.254***	
	(-2.267, -0.538)	(-6.525, 2.122)	(-0.156, 0.998)	(-0.424, -0.078)	114
TWICE × Quiz Score		0.130		-0.084	
		(-0.879, 1.138)		(-0.203, 0.031)	
PAPER	-1.471***	8.269	1.320***	0.056	
	(-2.331, -0.612)	(-4.351, 20.889)	(0.922, 1.718)	(-0.288, 0.236)	116
PAPER × Quiz Score		-1.652		-0.284***	
		(-4.001, 0.698)		(-0.389, -0.182)	
ENHANCED	-1.233***	-2.724	0.489*	-0.216**	
	(-2.083, -0.383)	(-6.883, 1.434)	(-0.030, 1.008)	(-0.402, -0.033)	113
ENHANCED × Quiz Score		0.345		-0.101*	
		(-0.560, 1.249)		(-0.212, 0.007)	
Quiz Score		-1.344***			
		(-1.872, -0.816)			
Intercept	0.373	6.236***	4.105***		
	(-0.090, 0.835)	(3.637, 8.836)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

Table B.5: Treatment effects on NMB3 and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB3		Quiz Score	NMB3	
	(1)	(2)	(3)	(4)	
NO QUIZ	0.223				n
	(-0.540, 0.987)				
ANSWERS	-0.442	-0.360	-0.050	-0.112	
	(-1.252, 0.369)	(-3.559, 2.839)	(-0.586, 0.487)	(-0.278, 0.070)	112
ANSWERS × Quiz Score		-0.075		0.009	
		(-0.851, 0.702)		(-0.089, 0.106)	
INCENTIVE	-0.759*	-2.207	0.211	-0.157*	
	(-1.810, 0.022)	(-5.334, 0.921)	(-0.354, 0.775)	(-0.336, 0.028)	114
INCENTIVE × Quiz Score		0.127		-0.037	
		(-0.508, 0.762)		(-0.141, 0.063)	
TWICE	-1.135***	-0.365	0.421	-0.183**	
	(-1.996, -0.274)	(-4.401, 3.670)	(-0.156, 0.998)	(-0.356, -0.004)	114
TWICE × Quiz Score		-0.193		-0.074	
		(-1.163, 0.776)		(-0.181, 0.026)	
PAPER	-1.342***	5.617	1.320***	0.074	
	(-2.220, -0.464)	(-2.618, 13.852)	(0.922, 1.718)	(-0.252, 0.278)	116
PAPER × Quiz Score		-1.143		-0.240***	
		(-2.649, -0.363)		(-0.340, -0.143)	
ENHANCED	-1.099**	0.321	0.489*	-0.158*	
	(-1.962, -0.235)	(-3.790, 4.431)	(-0.030, 1.008)	(-0.349, 0.028)	113
ENHANCED × Quiz Score		-0.314		-0.088*	
		(-1.201, 0.574)		(-0.189, 0.006)	
Quiz Score		-1.021***			
		(-1.471, -0.571)			
Intercept	0.105	4.400***	4.105***		
	(-0.350, 0.561)	(2.314, 6.486)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

Table B.6: Treatment effects on NMB – linear probability model robustness checks

	Dependent variable					
	NMB1		NMB2		NMB3	
NO QUIZ	-0.069		-0.011		0.055	
	(0.092)		(0.095)		(0.096)	
ANSWERS	-0.173*	-0.098	-0.092	-0.098	-0.110	-0.056
	(0.090)	(0.289)	(0.102)	(0.156)	(0.101)	(0.183)
INCENTIVE	-0.184**	-0.366	-0.197**	-0.316	-0.184*	-0.374
	(0.088)	(0.294)	(0.098)	(0.203)	(0.097)	(0.234)
TWICE	-0.237***	-0.283	-0.329 ***	-0.378*	-0.263***	-0.242
	(0.082)	(0.303)	(0.092)	(0.194)	(0.093)	(0.205)
PAPER	-0.220 ***	0.893*	-0.342***	0.911**	-0.301***	0.697
	(0.083)	(0.454)	(0.090)	(0.409)	(0.088)	(0.460)
ENHANCED	-0.206**	-0.126	-0.295***	-0.330	-0.256***	-0.111
	(0.086)	(0.347)	(0.095)	(0.254)	(0.094)	(0.250)
Quiz Score		-0.117***		-0.209***		-0.192***
		(0.035)		(0.022)		(0.026)
ANSWERS × Quiz Score		-0.020		-0.001		-0.016
		(0.060)		(0.040)		(0.043)
INCENTIVE × Quiz Score		0.048		0.038		0.053
		(0.059)		(0.042)		(0.048)
TWICE × Quiz Score		0.021		0.030		0.013
		(0.058)		(0.039)		(0.041)
PAPER × Quiz Score		-0.177**		-0.180**		-0.137*
		(0.079)		(0.069)		(0.078)
ENHANCED × Quiz Score		-0.005		0.030		-0.011
		(0.067)		(0.050)		(0.046)
Intercept	0.395***	0.873***	0.592***	1.450***	0.526***	1.313***
	(0.057)	(0.166)	(0.057)	(0.090)	(0.058)	(0.113)
Observations	308	265	308	265	308	265
R^2	0.044	0.194	0.082	0.387	0.072	0.340

QUIZ is the omitted category. *, **, and *** respectively denote $p < 0.1$, $p < .05$, $p < .01$.

Robust (HC1) standard errors are in parentheses.

across these treatments),[1] while 47.4% of subjects in the QUIZ treatment misunderstand. A simple ex-post power calculation indicates that if we recruited $n_1 = 76$ and $n_2 = 38$ subjects to two treatments in which each subject misunderstands with probability $p_1 = .474$ and $p_2 = .216$ (respectively), then we have a 79.4% chance of detecting a statistically significant difference between treatments (at the 5% significance level). This suggests a reasonable level of power in our comparisons between the four aforementioned treatments and QUIZ. However, 33.3% of subject misunderstand in the ANSWERS treatment – an intermediate case between QUIZ and these other four treatments. If we recruited $n_1 = 76$ and $n_2 = 36$ subjects to two treatments in which each subject misunderstands with probability $p_1 = .474$ and $p_2 = .333$ (respectively), then we have only a 33.2% chance of detecting a statistically significant difference between treatments. If instead we recruited $n_1 = 38$ and $n_2 = 36$ subjects to two treatments in which each subject misunderstands with probability $p_1 = .216$ and $p_2 = .333$ (respectively), then we have only a 18.2% chance of detecting a statistically significant difference between treatments. These calculations indicate that our sample sizes are too small to reliably detect a statistically significant difference between our ANSWERS treatment and the QUIZ treatment, or between the ANSWERS treatment and any of the INCENTIVE, TWICE, PAPER, and ENHANCED treatments. If we instead view the NO QUIZ and QUIZ, pooled, as baseline instructions treatments without reinforcement, and the remaining treatments as enhanced instructions or reinforcement treatments, then our samples have $n_1 = 119$, $n_2 = 189$, $p_1 = .462$, and $p_2 = .238$; under these samples sizes and NMB probabilities, we had a 98.3% chance of detecting a significant difference in NMB.

Our statistical analysis was conducted in R (R Core Team, 2017). The regressions in Table 2.4 (and above) used the 'lm' and 'glm' command in the base 'stats' package, with robust standard errors calculated using the 'sandwich' package (Zeileis 2004; 2006). Mediation analysis used the 'mediation' package (Tingley *et al.*, 2014). Goodman-Kruskal gamma tests use the 'DescTools' package (Signorell, 2018). We used the 'pwr' package (Champely, 2018) for the power analysis reported above. Figures made in 'ggplot2' (Wickham, 2009).

D Post-experiment questionnaire

At suggestion of a referee and the editor, we added a post-experiment questionnaire to our ENHANCED treatment, and ran additional sessions of the QUIZ treatment followed by this questionnaire to paint a more complete picture of subjects' decisionmaking processes as they went though the experiment. We asked nine questions in total.

Our first observation is that there is no statistical difference between QUIZ and ENHANCED on any of the first six quantitative questions.

[1]These numbers are relatively close to each other, so we use the 21.6% for our illustrative calculations below.

Figure B.21: Post-experiment questionnaire (Page 1)

Post-Experiment Questionnaire

Q1. Please think back to when you read the instructions and rate how much you agree with the following three statements on a scale of 1 to 7:

i. The instructions were clear.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

ii. I understood the best time to switch to task 2 (the slider task) – that is, when to switch in order to get the highest payment.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

iii. I understood that I could only complete task 2 once.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

Figure B.22: Post-experiment questionnaire (Page 2)

Q2. Please think back to when the experiment was underway and rate how much you agree with the following three statements on a scale of 1 to 7:

i. My main goal in the experiment was to maximize my earnings.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

ii. I remembered the best time to switch to task 2.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

iii. I remembered that I could only complete task 2 once.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

Figure B.23: Post-experiment questionnaire (Page 3)

Q3. Describe, in your own words, the rules of the experiment.

Q4. Describe, in your own words, how you decided whether and when to switch to task 2.

Q5. What advice would you give to a future participant in this experiment?

	QUIZ	ENHANCED	p-value
Comprehension			
Q1i (Clarity)	5.7 (6)	5.4 (6)	0.31
Q1ii (Understood Optimum)	5.7 (7)	5.6 (7)	0.41
Q1iii (Understood Once)	5.4 (7)	5.9 (7)	0.55
Retention			
Q2i (Maximized Earnings)	6.4 (7)	6.3 (7)	0.43
Q2ii (Remembered Optimum)	5.8 (7)	5.6 (6)	0.57
Q2iii (Remembered Once)	5.6 (7)	6.0 (7)	0.38

Mean (median) reported; p-values for rank-sum tests of equality of distributions.

Table B.7: Correlation between subjects' evaluation and misunderstanding and quiz score

	misunderstanding	p.value_misunderstanding	quiz score	p.value_score
Q1i	-0.168	0.159	0.281	0.017
Q1ii	-0.267	0.024	0.202	0.089
Q1iii	-0.406	0.0004	0.202	0.088
Q2i	0.039	0.744	0.046	0.700
Q2ii	-0.371	0.001	0.383	0.001
Q2iii	-0.356	0.002	0.196	0.100

Table B.7 shows that our post-experimental questionnaire results indicate that subjects largely felt that they both understood and retained the key pieces of information from the instructions – with the median subject indicating that they agreed or strongly agreed that they understood and remembered when they should switch (Q1ii, Q2ii), and how many times they could switch (Q1iii, Q2iii). In addition, most subjects agreed with the statement "The instructions were clear", with the median subject rating the statement a 6 out of 7. We find no significant differences between the distribution of answers to any of these questions between the QUIZ and ENHANCED treatments ($p > .3$ in all pairwise comparisons, rank-sum tests). Since we do observe a difference in NMB revealed in the experiment, our post-experimental questionnaire inadvertently reveals its limits at diagnosing reasons for NMB and the potential for improvements. That being said, Table C.3 indicates that subjects' post-experiment answers strongly correlate with both NMB in the experiment and quiz scores. Post-experiment reports of understanding (Q1ii,iii) and retention (Q2ii,iii) were each negatively correlated with NMB ($p < .03$ in all cases). In addition, the subject's post-experimental agreement with the statement "The instructions were clear" was positively correlated with their post-instructions quiz score ($\rho = .281$, $p = .017$).

22 of the 72 subjects who wrote the questionnaire mentioned the instructions in their written answers. Nearly all of these were in Q5: "What advice would you give to a future participant in this experiment?" For instance, the first three subjects to mention the instructions answered Q5 as follows: "Pay attention to the instructions." "Do the experiment with patience and read instructions very carefully." "Read the instructions and follow them for more $." These are typical answers; many subjects recognized, ex post, that paying close attention to the instructions was important for achieving the maximum payoff.

21 of the 72 subjects who wrote the questionnaire showed some kind of mistaken understanding of the experiment, even after having completed it. Many of these misunderstandings were orthogonal to our variable of interest (the time to do task 2). For instance, although our instructions clearly stated that one could get a $0.25 payoff for each period of task 1 if a certain threshold was reached, many seemed to believe that one could earn more than $0.25 by doubling or tripling the threshold. For instance, one subject wrote, "You have a poodle that jumps on to platforms, each 75 units, you get paid 25c." Another one wrote, "Roughly, I would only get 50c at most doing poodle jump for the whole period." The payoff is fixed at 25 cents, so 50 would be impossible. Many subjects appear to believe that they could earn for both tasks 1 and 2 if they completed the minimum height before switching. This is a minor misunderstanding, though it is stated in the instructions that one must forego earnings from one period of task 1 in order to perform task 2.

However, the majority of subjects do not show explicit misunderstandings in their answers, and some even demonstrate learning. One subject who did not perform task 2 at the correct time wrote, "I wasn't aware I can only switch to task 2 only once. So I switched to task 2 in the first period." Another wrote, "I thought it didn't mention number of times we could do the bonus so I did it very early on." These subjects clearly realized their mistakes after they had made them, which suggests that repeated decisions (with feedback of some form) can be a substitute for reinforcing understanding. On the other hand, some subjects failed to understand our instructions and still didn't understand them afterwards. One such subject wrote, "If you taking task 1, you can change game into task 2, but you cannot turn back to task 1."

B.4 Bibliography

Altmann, S., Falk, A., Grunewald, A. and Huffman, D. (2014). Contractual incompleteness, unemployment, and labour market segmentation. *The Review of Economic Studies*, **81** (1), 30–56.

Anderson, L. R., DiTraglia, F. J. and Gerlach, J. R. (2011). Measuring altruism in a public goods experiment: a comparison of US and Czech subjects. *Experimental Economics*, **14** (3), 426–437.

Aycinena, D., Baltaduonis, R. and Rentschler, L. (2014). Valuation structure in first-price and least-revenue auctions: an experimental investigation. *Experimental Economics*, **17** (1), 100–128.

Bayer, R.-C., Renner, E. and Sausgruber, R. (2013). Confusion and learning in the voluntary contributions game. *Experimental Economics*, **16** (4), 478–496.

Brookins, P. and Ryvkin, D. (2014). An experimental study of bidding in contests of incomplete information. *Experimental Economics*, **17** (2), 245–261.

Cabrera, S., Fatás, E., Lacomba, J. A. and Neugebauer, T. (2013). Splitting leagues: promotion and demotion in contribution-based regrouping experiments. *Experimental Economics*, **16** (3), 426–441.

Champely, S. (2018). *pwr: Basic Functions for Power Analysis*. R package version 1.2-2.

Cox, J. C. and James, D. (2012). Clocks and trees: Isomorphic dutch auctions and centipede games. *Econometrica*, **80** (2), 883–903.

Ericson, K. M. M. and Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *Quarterly Journal of Economics*, **126** (4), 1879–1907.

Etang, A., Fielding, D. and Knowles, S. (2011). Does trust extend beyond the village? experimental trust and social distance in cameroon. *Experimental Economics*, **14** (1), 15–35.

Harris, D., Herrmann, B., Kontoleon, A. and Newton, J. (2015). Is it a norm to favour your own group? *Experimental Economics*, **18** (3), 491–521.

Kamei, K., Putterman, L. and Tyran, J.-R. (2015). State or nature? endogenous formal versus informal sanctions in the voluntary provision of public goods. *Experimental Economics*, **18** (1), 38–65.

Mittone, L. and Ploner, M. (2011). Peer pressure, social spillovers, and reciprocity: an experimental analysis. *Experimental Economics*, **14** (2), 203–222.

Noussair, C. N. and Stoop, J. (2015). Time as a medium of reward in three social preference experiments. *Experimental Economics*, **18** (3), 442–456.

Petersen, L. and Winn, A. (2014). Does money illusion matter? comment. *American Economic Review*, **104** (3), 1047–62.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Signorell, A. (2018). *DescTools: Tools for Descriptive Statistics*. R package version 0.99.25.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, **59** (5), 1–38.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, **11** (10), 1–17.

— (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, **16** (9), 1–16.

Appendix C

Supplementary Appendix to "Mobility as a Service Apps and Multimodal Transportation: Evidence From a Multimodal App"

C.1 Robustness

In the main text of this paper, we chose to define a multimodal trip as one that started or ended within 1 meter of a transit stop. The choice of 1 meter was somewhat arbitrary. Using a larger radius creates the possibility of more false positives classified as multimodal while reducing false negatives. With a radius of 10 meters, the criteria is likely to include trips to businesses near bus stops. We do not expect these trips to be affected by the treatment, so it ultimately adds noise to the regression. Thus, re-running the analysis with a higher radius results in similar, but less statistically significant results. We present these results in Tables C.1 and C.2.

Table C.1: Effects of multimodal trip planning

Panel A: Dependent Variable: Whether a trip's origin or destination is within 10m of a bus stop						
	Logit Model			Probit Model		
	1	2	3	4	5	6
Multimodal trip planning	0.593	0.637	0.207	0.337	0.360	0.125
	(0.129)***	(0.122)***	(0.109)*	(0.072)***	(0.069)***	(0.064)**
	[0.108]***	[0.109]***	[0.114]*	[0.062]***	[0.062]***	[0.065]*
Month Fixed effects	N	Y	Y	N	Y	Y
Area Fixed effects	N	N	Y	N	N	Y
Panel B: Dependent Variable: Whether a trip's origin or destination is within 5m of a bus stop						
	Logit Model			Probit Model		
	1	2	3	4	5	6
Multimodal trip planning	0.482	0.544	0.247	0.259	0.286	0.138
	(0.185)***	(0.180)***	(0.164)	(0.095)***	(0.093)***	(0.087)
	[0.135]***	[0.137]***	[0.141]*	[0.071]***	[0.071]***	[0.075]*
Month Fixed effects	N	Y	Y	N	Y	Y
Area Fixed effects	N	N	Y	N	N	Y

Notes: The dependent variable is an indicator whether variable showing whether a trip is multimodal, i.e., its origin or destination is within 10m/5m of a bus stop. Standard errors clustered at the month level are in parentheses and Newey-West standard errors are in brackets. All the trips' origin or destination in the sample is within 5km of a bus stop. Cities with less than 50 trips are excluded.

* Statistical significance at the 0.1 level.

** Statistical significance at the 0.05 level.

*** Statistical significance at the 0.01 level.

Table C.2: Effects of multimodal trip planning on existing users

Panel A: Dependent Variable: Whether a trip's origin or destination is within 10m of a bus stop						
	Logit Model			Probit Model		
	1	2	3	4	5	6
Multimodal trip planning	0.491 (0.154)*** [0.123]***	0.535 (0.145)*** [0.124]***	0.060 (0.134) [0.142]	0.277 (0.088)*** [0.071]***	0.298 (0.083)*** [0.071]***	0.040 (0.079) [0.08]
Month Fixed effects	N	Y	Y	N	Y	Y
Area Fixed effects	N	N	Y	N	N	Y
Panel B: Dependent Variable: Whether a trip's origin or destination is within 5m of a bus stop						
	Logit Model			Probit Model		
	1	2	3	4	5	6
Multimodal trip planning	0.291 (0.194) [0.163]	0.355 (0.187)* [0.164]**	-0.030 (0.197) [0.198]	0.152 (0.101) [0.085]	0.179 (0.098)* [0.086]**	-0.006 (0.104) [0.098]
Month Fixed effects	N	Y	Y	N	Y	Y
Area Fixed effects	N	N	Y	N	N	Y

Notes: The dependent variable is an indicator whether variable showing whether a trip is multimodal, i.e., its origin or destination is within 10m/5m of a bus stop. Standard errors clustered at the month level are in parentheses and Newey-West standard errors are in brackets. All the trips' origin or destination in the sample is within 5km of a bus stop. Cities with less than 50 trips are excluded. Only existing users' data is used. Existing users are users who first installed the app prior to the introduction of the multimodal feature.

* Statistical significance at the 0.1 level.

** Statistical significance at the 0.05 level.

*** Statistical significance at the 0.01 level.